PolyPublie
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Détection et suivi du visage et des mains appliqués à la surveillance de prise de médicaments
Auteur: Author:	Soufiane Ammouri
Date:	2008
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ammouri, S. (2008). Détection et suivi du visage et des mains appliqués à la surveillance de prise de médicaments [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/8314/

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/8314/
Directeurs de recherche: Advisors:	
Programme: Program:	Non spécifié

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

UNIVERSITÉ DE MONTRÉAL

DÉTECTION ET SUIVI DU VISAGE ET DES MAINS APPLIQUÉS À LA SURVEILLANCE DE PRISE DE MÉDICAMENTS

SOUFIANE AMMOURI
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INFORMATIQUE)
AOUT 2008

© Soufiane Ammouri, 2008.

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-46028-3
Our file Notre référence
ISBN: 978-0-494-46028-3

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

CE MÉMOIRE INTITULÉ :

DÉTECTION ET SUIVI DU VISAGE ET DES MAINS APPLIQUÉS À LA SURVEILLANCE DE PRISE DE MÉDICAMENTS

présenté par : AMMOURI Soufiane

en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées

a été dûment accepté par le jury d'examen constitué de :

M. LANGLOIS J.M. Pierre, Ph.D., président

M. BILODEAU Guillaume-Alexandre, Ph.D., membre et directeur de recherche

M. OZELL Benoit, Ph.D., membre

À la recherche objective.

À la paix dans le monde.

REMERCIEMENTS

J'aimerais remercier les membres du jury pour avoir accepté d'évaluer mon mémoire.
Je souhaite remercier chaleureusement le directeur de mon projet Monsieur Guillaume-Alexandre Bilodeau, pour m'avoir accueilli dans son laboratoire, la confiance qu'il a faite en moi, ses aimables et valeureuses directives pour mener à bien ce projet et surtout sa totale disponibilité pour répondre à mes questions.
Je veux aussi remercier Chantal Balthazard et Jeanne Daunais, respectivement commis et secrétaire au département de génie informatique et génie logiciel de l'École Polytechnique de Montréal, pour les différents services qu'elles rendent aux étudiants.
Par la suite, je tiens à remercier tous les membres de LITIV, je vous dis tout simplement que travailler avec vous était une expérience très agréable pour moi.
Je ne veux pas oublier de remercier toutes les personnes qui ont pris le temps d'aller avec moi pour les prendre en photo ou en vidéo afin de construire la base de données sur laquelle j'ai travaillé et j'ai testé mes résultats. Sans eux, je serais sûrement incapable de réaliser ce projet.
Je tiens à remercier particulièrement James Hudon (stagiaire au LITIV) qui m'a aidé à comparer mes méthodes de suivi du visage et des mains avec d'autres méthodes existantes.
Finalement, j'aimerai bien remercier les membres de ma famille ainsi que tous ceux et celles qui m'ont soutenus tout au long du déroulement de ma maitrise, tant moralement que techniquement. Ils étaient ma principale source de motivation et d'encouragement. Plus particulièrement, je tiens à remercier mes parents, les êtres les plus chers dans ma vie.

RÉSUMÉ

Ce mémoire présente un système de détection et de suivi des parties du corps dans un flux vidéo. Les techniques de localisation et du suivi du visage et des mains sont utilisées par la suite pour la détection d'activités humaines. Dans ce travail, on a choisit la détection automatique de prise de médicaments dans des séquences vidéo prises par une caméra statique. Le travail s'effectue sans la reconnaissance de la position exacte du comprimé à cause de sa petite taille qui empêche sa localisation et son suivi. La détection des parties du corps de la personne et des bouteilles de médicaments est effectuée à l'aide de techniques basées sur la couleur et la forme. Pour le suivi de ces objets, on utilise plusieurs méthodes basées sur les histogrammes de couleurs, les moments de Hu et les contours. Pour la reconnaissance de la prise de médicaments, on se base sur un réseau de Petri afin de s'assurer que les états permettant la détection de prise de médicaments sont parcourus dans le bon ordre dans la séquence à analyser. Dans des conditions contrôlées, nos algorithmes de détection et de suivi du visage et des mains ont une efficacité de plus de 96% et permettent la détection de la prise de médicaments dans différents scénarios.

Mots clef – Détection du visage et des mains, suivi du visage, suivi des mains, médicaments, Moments de Hu, histogrammes de couleurs, contours, réseau de Petri.

ABSTRACT

This thesis presents detection and tracking methods for user's body parts in video sequences. They are applied to detect human activities. In this work, we chose the automatic detection of medication intake in video taken by a static camera. The work is done without the recognition of the location of pills because of its small size, which prevents locating and monitoring. We use a technique based on color and shape to detect the body parts and the medication bottles. Color is used for skin detection, and the shape is used to distinguish the faces from the hands and differentiate bottles of medicine. To track these objects, we use methods based on color histograms, Hu moments and edges. For the recognition of medication intake, we use a Petri network and event recognition. In controlled conditions, our methods have an accuracy of more than 96% and allow the detection of the medication intake in various scenarios.

Keywords - Face and hands detection, face tracking, hand tracking, medication, Hu moments, color histograms, edges, Petri network.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

Blob	Binary Large OBject (groupe de pixels)
EF	Eigenfaces
EO	Eigenobjects
HSV	Hue (teinte) Saturation Valeur
RGB	Red Green Bleu (rouge vert bleu)
RN	Réseaux de neurones
HMM	Hidden Markov model (modèle de Markov caché)
MS	Merge-split (Fusion-Séparation)
ST	Straight-through
MDPA	Minimum distance of pair assignements
OpenCV	Open Source Computer Vision

LISTE DES ANNEXES

INTRODUCTION

La vidéosurveillance consiste à placer des caméras de surveillance dans un lieu public ou privé pour pouvoir visualiser et analyser ce qui s'y passe. Les applications en vidéosurveillance sont aussi nombreuses que diversifiées. En matière de sécurité, elle est utilisée pour surveiller les allers et venues, prévenir les vols, agressions, fraudes et gérer les incidents et mouvements de foule. Dans le domaine biomédical, la vidéosurveillance peut être utilisée pour la détection de chutes, l'analyse d'habitudes alimentaires ou encore pour le contrôle de la prise de médicaments. Dans ce mémoire, il sera question de cette dernière application.

Contexte et problématique

« *Les personnes de plus de 85 ans sont un peu plus d'un million et dans dix ans, elles seront près du double* », a indiqué en 2006 Philippe Bas, ministre délégué aux Personnes âgées de la France, lors des 13èmes rencontres parlementaires tenues à Paris sur la longévité ayant pour thème « Longévité et nouvelles technologies » [1]. Aussi, une étude de *l'Agence de santé publique de Canada* [2] estime qu'en 2016, la proportion de Canadiens de 65 ans et plus dans la population sera de 16% et passera à plus de 22% en 2041. En effet, le Canada assiste actuellement à un vieillissement important de sa population. L'augmentation de l'âge augmente les problèmes de santé et entraine par conséquent une augmentation de la prise des médicaments. Sachant que la technologie peut améliorer de manière décisive la qualité de vie des personnes âgées ou ayant une déficience mentale en leur permettant de rester chez elles plus longtemps et en leur offrant un suivi médical à la fois plus souple et plus efficace, on s'est intéressé au problème du contrôle de la prise de médicaments. Ainsi le domaine de la vidéosurveillance a suscité beaucoup d'intérêt mais il n'a été que peu dirigé vers le contrôle de la prise de médicaments.

Récemment, quelques articles ([3], [4]) ont traité la détection de prise de médicaments. Les méthodes utilisées dans ces deux articles seront bien décrites et analysées au chapitre 1. Pour détecter l'activité humaine qui est dans notre cas la prise de médicaments, on doit

pouvoir localiser et suivre les objets qui interagissent dans cette activité. Ces objets sont le visage, les mains et les bouteilles de médicaments. Quelque soit l'objet que l'on veut suivre dans une séquence vidéo, on a besoin d'un modèle pour le décrire : ce modèle peut contenir aussi bien de l'information a priori sur l'objet que de l'information extraite des trames précédentes. Il peut être constitué de descripteur de couleur, de texture, de forme ou de tout autre type de primitives. Les différentes méthodes développées pour la détection et le suivi de ces objets seront décrites et analysées au chapitre 1.

Objectifs

Un système complet pour contrôler la prise de médicaments doit :

1. Identifier la personne qui est en train de prendre le médicament.
2. Détecter la prise du médicament.
3. Identifier le médicament pris.
4. Détecter la quantité prise de ce médicament.
5. Détecter l'heure de prise du médicament.

Notre système se concentre sur les problèmes (2) et (3). Le problème (1) sera parmi les travaux futurs applicables à notre projet, le problème (4) est difficile à traiter si on ne peut localiser et suivre les comprimés puisqu'ils sont toujours occultés par les doigts, et le dernier problème est relativement facile à résoudre mais il ne sera pas discuté dans ce mémoire. Donc, le but de la recherche est de développer des méthodes pour :

- Détecter et suivre le visage et les mains dans une séquence vidéo.
- Localiser et identifier les bouteilles de médicaments présentes dans la séquence.
- Détecter la prise de médicament en identifiant le médicament pris.

Aperçu de la méthode proposée

On commence par filmer une scène dans laquelle un usager prend ses médicaments que nous présentons à notre système. Pour chacune des images (trames) de la séquence,

l'espace de couleur *HSV* avec des seuils prédéfinis est utilisé pour détecter les régions de peau contenues dans cette séquence. Des hypothèses sur la forme du visage nous permettent de localiser ce dernier dans la trame initiale. Le suivi du visage dans les trames suivantes s'effectue en utilisant un descripteur de forme. Le suivi des mains s'effectue en exploitant les propriétés de contours. L'identification et la localisation des bouteilles de médicament se fait en combinant les histogrammes de couleurs et un descripteur de forme et le suivi de ces derniers se base sur la propriété du centroïde.

Nos algorithmes de localisation et de suivi des parties du corps et des bouteilles de médicaments sont appliqués pour la détection de l'activité humaine. Dans notre cas, on a utilisé un réseau de Petri afin de reconnaître la prise de médicament en définissant différents états liés à l'action de prise de médicaments.

Contributions

Notre contribution est essentiellement la réalisation d'un nouveau système de détection de l'activité humaine qui est dans notre cas la prise de médicaments. Spécifiquement, notre première contribution est la création d'un nouvel algorithme de suivi du visage se basant sur la détection des régions de la peau, sur quelques hypothèses sur la forme du visage pour pouvoir le localiser dans la trame initiale, et sur les moments de Hu qui seront expliqués au chapitre 2 pour effectuer le suivi. Comme deuxième contribution, on a utilisé la segmentation en région de peau et exploité les propriétés de contours, de morphologie mathématique (dilatation), de la densité des arêtes ainsi que du centroïde des régions pour pouvoir localiser les mains dans les bras et pouvoir les suivre. Notre troisième contribution est la proposition d'une méthode qui combine les histogrammes de couleurs et les moments de Hu pour l'identification des bouteilles de médicaments. Enfin, comme quatrième contribution, on a conçu un réseau de Petri afin de reconnaitre la prise de médicament. Dans ce réseau, la transition des jetons d'une place à l'autre ne se produit que si les événements durent un certain nombre de trames.

Plan du mémoire

À travers ce mémoire, on va discuter des méthodes utilisées ainsi que des résultats obtenus afin d'atteindre les objectifs du projet. Le reste de ce document est structuré comme suit. Le premier chapitre présente une brève revue de l'état de l'art. Ce dernier décrit les méthodes de détection de la peau humaine, les méthodes de détection du visage et des mains, les différentes approches du suivi des objets, les techniques déjà utilisées pour la reconnaissance de l'activité humaine et deux articles portant directement sur des systèmes de contrôle de prise de médicaments. Le second chapitre décrit en détails notre méthodologie ou contribution, c'est-à-dire les algorithmes qui sont implémentés dans le cadre de ce mémoire. On y trouve la méthode utilisée pour la détection des régions de la peau contenues dans chacune des trames de la séquence vidéo, la technique adoptée pour la détection et la gestion des occlusions entre les parties du corps suivies, les méthodes utilisées pour la détection et le suivi du visage et des mains, les techniques de localisation et d'identification des bouteilles de médicaments et le réseau de Petri construit pour la détection de l'activité humaine. Finalement, dans le troisième chapitre, on fournit les résultats et les performances de nos algorithmes ainsi qu'une brève discussion.

CHAPITRE 1 REVUE DE LA LITTÉRATURE

Sachant que la technologie peut améliorer de manière décisive la qualité de vie des personnes âgées ou ayant une déficience mentale en leur permettant de rester chez elles plus longtemps et en leur offrant un suivi médical à la fois plus souple et plus efficace, on a choisi d'appliquer nos méthodes pour le contrôle de la prise de médicaments. Ainsi le domaine de la vidéosurveillance a suscité beaucoup d'intérêt mais il a été que peu dirigé vers le contrôle de la prise de médicaments. Récemment, quelques articles ([3], [4]) ont traité la détection de prise de médicaments. Ces articles seront analysés dans un premier temps. Par la suite, on va discuter des méthodes qui traitent la détection et le suivi de même que la reconnaissance des activités humaine.

1.1 Surveillance de la prise de médicament

Dans cette section, on va présenter les méthodes utilisées dans les articles [3] et [4] dans lesquelles les auteurs ont présenté un système de vision par ordinateur permettant la surveillance du comportement de prise de médicaments.

1.1.1 Article de Batz et al. 2005 [3]

1.1.1.1 Description de la méthode

Dans leur approche, les auteurs commencent par détecter les régions de la peau contenues dans chacune des trames. Pour ce faire, ils extraient manuellement les pixels de peau des images de trois personnes différentes. Les pixels sont transformés par la suite à l'espace de couleur *YCbCr* pour rendre leur chrominance (couleur) plus indépendantes à leur luminance (intensité). Des intervalles de seuillage sont construits à l'aide des valeurs des pixels extraits afin d'effectuer la binarisation de chacune des images (1: peau, 0: non-peau). La segmentation se fait en utilisant un algorithme de composantes connectées suivi des opérations morphologiques telles qu'un filtre médian. Une fois les régions de peau étiquetées, les auteurs se basent sur les rectangles englobant et les centroïdes de ces régions pour détecter la présence des occlusions entre les mains et le visage. La

dimension et la forme des régions permettent de différencier les mains du visage. La bouche est localisée selon la couleur des lèvres. Les bouteilles de médicaments sont détectées en cherchant dans l'image de contours des objets de formes rectangulaires et de proportion hauteur/largeur d'approximativement 2 :1. Une librairie de bouteilles définie permet la localisation d'objets dans les régions respectant ces proportions. Un appariement de gabarits permet le suivi de ces bouteilles en effectuant des corrélations sur les composantes *Cb* et *Cr* entre les régions possibles et le gabarit soumis à des rotations et des translations. Le système détecte la prise de médicaments si la séquence formée par les événements « Ouverture de la bouteille », « Main sur la bouche » et « Fermeture de la bouteille » se produit. L'ouverture et la fermeture des bouteilles sont détectées en analysant l'orientation des doigts dans les régions qui représentent les mains.

1.1.1.2 Analyse de la méthode

Dans cet article, la segmentation initiale de l'image en régions de peau se fait en se basant sur des pixels de peau extraits manuellement pour seulement trois personnes. Donc, il est fort possible que le système échoue dans la détection de la couleur de la peau en présence d'une nouvelle personne. Aussi, après la localisation du visage, les auteurs utilisent la transformation développée par [5] afin de faire ressortir les lèvres et détecter la bouche. Dans [5], Hsu *et al.* ont remarqué que la couleur de la région qui représente les lèvres contient une plus forte composante rouge et une faible composante bleu que les autres régions du visage. Par conséquent, la composante de chrominance *Cr* est supérieure à *Cb* dans les lèvres. Ils ont constaté que le rapport *Cr/Cb* est plus faible que la valeur de Cr^2 et par conséquent la différence Cr^2- (*Cr/Cb*) est beaucoup plus importante pour les lèvres que pour les autres parties du visage. L'image construite permettant la localisation de la bouche est définie avec les équations

$$Carte\_Bouche = C_r^{\,2}.(C_r^{\,2} - \eta.\frac{C_r}{C_b})^2, \tag{1.1}$$

où $$\eta = 0.95.\frac{\frac{1}{n}\sum_{(x,y)\in\Re} C_r(x,y)^2}{\frac{1}{n}.\sum_{(x,y)\in\Re} C_r(x,y)/C_b(x,y)}, \quad (1.2)$$

avec Cr^2 et Cr/Cb qui sont normalisées entre 0 et 255, n qui est le nombre de pixels de la région $\Re$ représentant le visage et (x,y) représente les coordonnées des pixels se trouvant dans la région $\Re$. Après quelques opérations morphologiques, ils seuillent cette carte pour localiser la bouche. Dans [6], Eveno *et al.* doutait de l'homogénéité de cette expression pour l'extraction des lèvres. Dans le cadre de notre étude, on a implémenté les équations 1.1 et 1.2 afin de tester leurs homogénéités et leurs performances pour la détection de la bouche. La figure 1.1 compile les résultats obtenus pour deux personnes différentes.

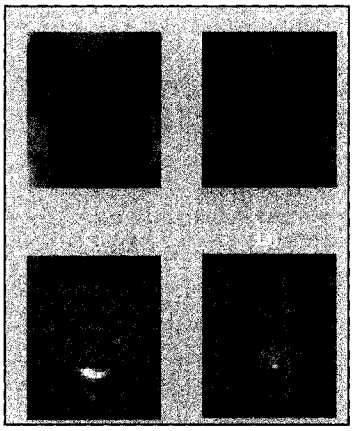

Figure 1.1 Localisation des lèvres avec la méthode de Hsu *et al.*[5] A), B) trames des séquences originales. C), D) Carte de la bouche permettant de localiser les lèvres.

On peut remarquer que les lèvres peuvent être détectées lorsque la personne porte du rouge à lèvres (figure 1.1A) ou si les lèvres de la personne possèdent une très forte composante rouge. Dans le cas contraire, la méthode utilisée ne permet pas la détection des lèvres (figure 1.1B) et par conséquent la localisation de la bouche.

De plus, comme mentionné précédemment, les auteurs se basent sur l'orientation des doigts afin de reconnaitre les actions d'ouverture et de fermeture de la bouteille de médicaments. Dans ce cas, la localisation et le suivi des mains doivent être parfaitement

précis et les doigts toujours visibles, ce qui n'est que rarement possible. Finalement, pour la détection de la prise de médicament, la personne peut toujours ouvrir la bouteille pour prendre la pilule et par la suite la fermer avant de mettre le comprimé dans la bouche. Dans ce cas, le système ne détectera pas la prise de médicament car les actions « Main sur la bouche » et « Fermeture de la bouteille » sont interverties. De plus, le fait de détecter la prise de médicaments si une séquence d'actions se produit à chaque trame sans analyser la durée de ces actions peut engendrer une augmentation du taux de fausses détections dans le système conçu.

1.1.2 Article de Valin et al. 2006 [4]

1.1.2.1 Description de la méthode

Dans leur approche, trois types d'objets mobiles sont détectés et suivis: la tête de la personne, ses mains et les bouteilles de médicaments. Pour la détection et le suivi de la tête, les auteurs ont utilisé l'algorithme présenté dans [7]. La tête est modélisée en une ellipse dont la taille peut varier d'une trame à l'autre. Pour chaque image, une recherche locale détermine l'ellipse qui représente le mieux la tête. Pour ce faire, l'algorithme se base sur l'intensité du gradient, sur le périmètre de l'ellipse, et sur la vraisemblance de la couleur de la peau à l'intérieur de celle-ci. Dans [7], les auteurs utilisent la notation s=(x, y, σ) qui correspond à une ellipse de centre (x, y) et de demi-petit axe de longueur σ, la meilleur ellipse représentant la tête s* vérifie l'équation

$$s^* = \arg\max_{s_i \in S} \{\phi_g(s_i) + \phi_c(s_i)\}, \tag{1.3}$$

où $\phi_g(s_i)$ et $\phi_c(s_i)$ représentent respectivement des scores de correspondance pour l'intensité du gradient et la vraisemblance de couleur de la peau et S correspond à l'ensemble des ellipses pouvant contenir la tête. L'intensité du gradient d'un pixel correspond au taux de changement de l'intensité dans une image en tons de gris et dans une petite région avoisinante. La carte de gradients de l'image est obtenue en appliquant un filtre Gaussien puis un filtre de Sobel à l'image en niveaux de gris. Le score du gradient est défini selon l'équation

$$\phi_g(s) = \frac{1}{N_\sigma} \sum_{i=1}^{N_\sigma} \left| n_\sigma(i).g_s(i) \right|, \tag{1.4}$$

où $g_s(i)$ représente l'intensité du gradient au pixel *i* du périmètre de l'ellipse s, N_σ est le nombre de pixels sur le périmètre d'une ellipse de demi-petit axe σ et $n_\sigma(i)$ est le vecteur normal à l'ellipse au pixel *i*. La carte de vraisemblance de couleur peau est obtenue en calculant l'intersection entre l'histogramme du modèle de couleur peau *M* et l'histogramme de l'image dans l'ellipse *I*. Le score de couleur pour chaque ellipse est obtenu selon l'équation

$$\phi_c(s) = \frac{\sum_{i=1}^{N} \min(I_s(i), M(i))}{\sum_{i=1}^{N} I_s(i)}. \tag{1.5}$$

L'espace de couleur utilisé combine les composantes *R*, *G* et *B*. Cet espace de couleur appelé *color123* et défini par les équations 1.6 contient deux premières composantes de chrominance et une dernière composante de luminosité.

$$\begin{aligned} color1 &= \max(0,\min(255,(B\text{-}G)*10+128)) \\ color2 &= \max(0,\min(255,(G\text{-}R)*10+128)) \\ color3 &= \max(0,\min(255,(R+G+B)/3)) \end{aligned} \tag{1.6}$$

Le positionnement et le suivi des mains sont effectués en déterminant les régions de couleur peau les plus susceptibles de représenter les mains. Les auteurs se basent sur la carte de vraisemblance de couleur peau créée dans l'étape précédente et sur un algorithme de composantes connexes afin d'extraire les régions de peau. Ils utilisent quelques hypothèses pour éliminer les régions qui ne sont pas susceptibles d'être les mains. Les occlusions possibles entre les deux mains et entre les mains et la tête sont détectées en se basant sur le nombre des régions de la peau obtenues et les positions précédentes des mains. Pour pouvoir différencier les bouteilles de médicaments, des bandes de couleur sont collées sur ces dernières. Pour la détection des bouteilles de médicaments, les auteurs ont utilisé le modèle de couleur décrit dans [8]. L'espace de

couleur utilisé dans ce cas est *YCbCr* et la vraisemblance de couleur est définie comme étant fonction de la distance de Mahalanobis *d* avec le modèle. Cette distance est définie par l'équation

$$d^2 = (x - \mu_m)^T \sum_m^{-1} (x - \mu_m), \tag{1.7}$$

où x est le vecteur couleur en trois dimensions du pixel et μ_m et $\sum_m^{-1}$ sont respectivement le vecteur moyen et l'inverse de la matrice de covariance de la distribution du modèle. Un ensemble d'images contenant des régions de chacune des bandes de couleurs utilisées est présenté au système afin de créer le modèle de couleur des bouteilles et en définir la distribution. Dans ce travail, la reconnaissance d'activités humaines est basée sur le concept de scénario. Ce dernier est une activité impliquant des objets mobiles et qui se déroule sur une certaine période de temps. L'algorithme développé est formé de trois niveaux de scénarios : état-simple, état-multiple et complexe. La figure 1.2 montre la relation entre les différents niveaux de scénarios.

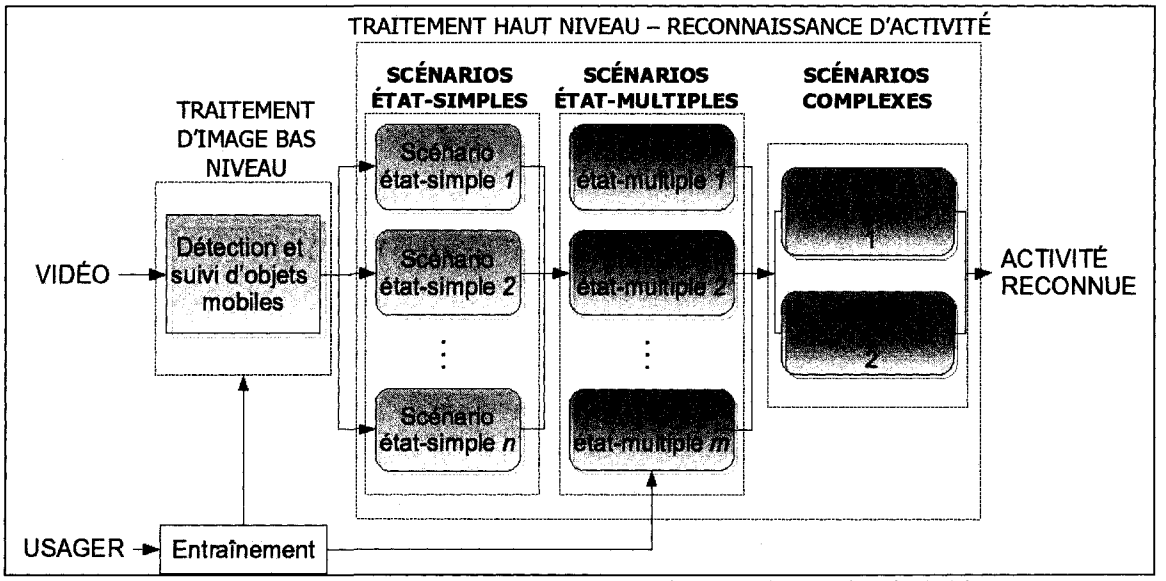

Figure 1.2 Structure du modèle de scénarios à trois niveaux. Figure adaptée de [4]

Les auteurs ont utilisé les concepts de scénarios états-simples et états-multiples élaborés dans [9]. Un scénario d'état simple est défini par un ensemble de caractéristiques d'objets

mobiles. Les scénarios états-simples élaborés par les auteurs pour la détection de la prise de médicaments sont définis comme suit :

- S_1 : Une seule main manipule la bouteille de médicaments,
- S_2 : Deux mains manipulent la bouteille de médicaments,
- S_3 : Une main touche à la tête,
- S_4 : Une main s'approche de la tête,
- S_5 : Une main s'éloigne de la tête.

Un scénario état-multiple correspond à une séquence de scénarios états-simple et il est évalué sur une longue période de temps. Trois scénarios états-multiples sont définis pour la reconnaissance de prise de médicaments :

- MS_1 : La personne ouvre la bouteille de médicaments et prend les pilules ($S_2 \rightarrow S_1 \rightarrow S_2 \rightarrow S_1$),
- MS_2 : La personne avale les pilules ($S_4 \rightarrow S_3 \rightarrow S_5$),
- MS_3 : La personne referme la bouteille de médicaments ($S_1 \rightarrow S_2 \rightarrow S_1$).

Afin de modéliser tous les cas possibles de prise de médicaments, les auteurs séparent l'activité complexe en séquence de scénarios états-multiples. Les deux représentations du scénario complexe sont présentées à la figure 1.3.

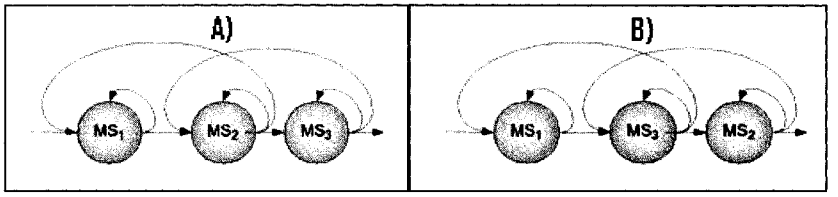

Figure 1.3 Systèmes représentant le scénario complexe constitué des séquences de scénarios états-multiples. A) Première représentation du scénario complexe {MS1, MS2, MS3}, B) Deuxième représentation du scénario complexe {MS1, MS3, MS2}. Figure adaptée de [4]

1.1.2.2 Analyse de la méthode

Premièrement, l'utilisation d'un modèle elliptique pour la détection et le suivi de la tête dans le cadre de l'activité de la prise de médicament n'est pas celui le plus approprié. En effet, ce genre de modèle se prête davantage à des activités où la tête de la personne reste toujours face à la caméra. Les interactions et les contacts entre les mains et

la tête peuvent aussi fausser le processus de suivi et par conséquent l'activité humaine ne sera pas correctement détectée.

Deuxièmement, pour la détection et le suivi des mains, plusieurs hypothèses simplificatrices du problème ont été utilisées. Parmi ces hypothèses, on trouve le fait de supposer que la personne qui prend les médicaments porte toujours un chandail à manches longues. Cette hypothèse est introduite pour éviter la localisation de la main dans le bras ce qui rend le système plus contrôlé.

Finalement, pour la détection et le suivi des bouteilles de médicaments, les auteurs utilisent des bandes de couleur qui sont collées aux bouteilles. Le fait de se baser uniquement sur la couleur peut augmenter les fausses détections du système surtout avec la présence dans l'environnement de prise de médicaments d'objets ayant des couleurs semblables à celles des bandes utilisées.

1.2 Détection et suivi

Dans cette section, on décrit les méthodes qui existent et qui traitent la détection et le suivi des objets. Pour le suivi des parties du corps telles que le visage et les mains, plusieurs auteurs détectent les régions de la peau en se basant sur la propriété de la couleur. Cette étape est très utilisée pour la localisation de ces parties et l'initialisation automatique de l'algorithme de suivi. On propose un aperçu des approches utilisant cette technique à la section 1.2.1. Les méthodes traitant la détection du visage et des mains sont décrites à la section 1.2.2. Finalement, la section 1.2.3 présente les méthodes qui existent et qui sont utilisées pour le suivi des objets.

1.2.1 Détection de la couleur de la peau

Pour pouvoir détecter les mains et le visage, plusieurs méthodes se basent sur la couleur de la peau afin de segmenter les images en régions de cette couleur. En effet, la peau humaine est souvent représentée par une portion d'un espace de couleur particulier et il est par conséquent possible d'extraire les pixels dont la couleur peut s'apparenter à celle de la peau. Il existe plusieurs espaces de couleurs s'appliquant à la détection de la peau.

Dans [10], les auteurs ont élaboré un classificateur de peau en définissant explicitement (par le biais d'un certain nombre de règles) les limites peau d'un pixel dans l'espace de couleur RGB. Dans leur approche, un pixel est classifié comme étant de la peau si $R > 95$ et $G > 40$ et $B > 20$ et $max\{R, G, B\}-min\{R, G, B\} > 15$ et $|R - G| > 15$ et $R > G$ et $R > B$. Dans [11], les auteurs utilisent l'espace de couleur RGB normalisé qui est obtenu en utilisant une simple normalisation comme l'expliquent les équations

$$\begin{aligned} r &= \frac{R}{R+G+B}, \\ g &= \frac{G}{R+G+B}, \\ b &= \frac{B}{R+G+B}. \end{aligned} \tag{1.8}$$

Les trois composants normalisés r, g et b sont appelées les couleurs pures puisqu'elles ne contiennent aucune information sur la luminance. La somme des trois composants est toujours égale à un, ainsi seulement deux composantes r et g sont employées pour décrire complètement l'espace de couleur représentant la peau. Pour établir le modèle, les auteurs ont rassemblé des échantillons de peau humaine de différentes couleurs (i.e. pour différentes races). Pour chaque pixel de peau prélevé, les valeurs de r et g sont calculées puis la moyenne (μ_r et μ_g) et l'écart type (σ_r et σ_g) de r et de g dans tous les échantillons de peau sont calculés. Leur détecteur de peau examine chaque pixel de l'image d'entrée et calcule ses valeurs r et g. Si ces valeurs satisfont les équations 1.9, alors ce pixel est considéré comme étant de la peau. La valeur de α détermine la précision du détecteur de peau et sa valeur est trouvée expérimentalement. La sortie du détecteur de peau est un masque binaire qui contient des uns dans les régions de peau et des zéros dans des régions de non-peau.

$$\begin{aligned} \mu_r - \alpha\sigma_r < r < \mu_r + \alpha\sigma_r \\ \mu_g - \alpha\sigma_g < g < \mu_g + \alpha\sigma_g \end{aligned} \tag{1.9}$$

Plusieurs auteurs utilisent des espaces de couleurs qui séparent la chrominance de la luminosité. En effet, plusieurs études [11, 12] ont montré que les types de peau diffèrent par la luminosité plutôt que par la chrominance. Il suffit de coder l'image dans

un système de couleur séparant l'intensité de la chrominance. Comme dans [3], des auteurs semblent préférer l'espace de couleur *YCbCr* pour la détection de la peau. Cet espace de couleur sépare la chrominance (*CbCr*) de la luminance (*Y*), ce qui permet une meilleure représentation de la couleur de la peau humaine. Des études antérieures [13, 14] proposent des seuils de détection des pixels appartenant à la peau dans l'espace de couleur *YCbCr*. La figure 1.4 extraite de l'étude [14] montre la zone et les seuils qui englobent tous les pixels de couleurs de la peau.

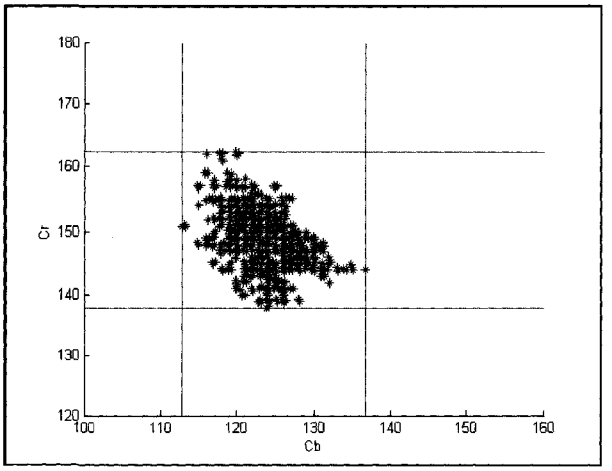

Figure 1.4 Zone des valeurs de *Cb* et *Cr* pour la détection de la couleur de la peau. Figure extraite de [14].

Dans [13], les auteurs transforment l'espace de couleur *RGB* à l'espace de couleur *YCbCr* en utilisant l'équation de transformation 1.10.

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112 \\ 112 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1.10}$$

Après la transformation, les auteurs ont généré une distribution de la couleur de la peau dans le plan *Cb-Cr* qui leur a permis de l'illustrer comme étant une fonction de probabilité conditionnelle. Par la suite, ils utilisent un classificateur Bayésien avec deux

classes pour pouvoir classifier les pixels de l'image d'entrée. La classe w_1 pour désigner les pixels de couleur peau et la classe w_2 pour les pixels de couleur non-peau.

Dans [15], les auteurs considèrent que l'utilisation des simples composantes de chrominance ne permet de représenter la couleur peau que sur une petite plage d'intensités lumineuses. Ainsi, le modèle de couleur proposé fait intervenir la composante de luminance dans la classification. Les seuils de décision devraient alors être différents selon la valeur du canal *Y* représentant la luminance comme le montre la figure 1.5. Les auteurs représentent la couleur peau par trois sous-modèles, un pour chaque plage de luminance. La classification s'effectue en comparent chaque pixel aux trois sous-modèles en utilisant un seuil et la distance de Mahalanobis présentée précédemment.

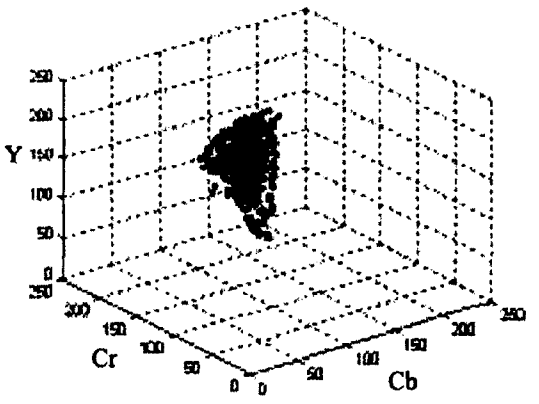

Figure 1.5 Représentation de la couleur peau dans l'espace *YCbCr*. Figure extraite de [15]

D'autres auteurs utilisent l'espace de couleur *HSV* pour la détection de la couleur de la peau. L'avantage du HSV pour la détection des couleurs réside dans le fait que le canal *V* (Value) représente la luminance permettant ainsi d'exprimer les couleurs sans se soucier des variations de luminosité. Dans [16], les auteurs extraient les pixels appartenant à la peau en observant seulement la teinte (*H*) et la saturation (*S*). Dans leur approche, un pixel est classifié comme étant de la peau si sa saturation est entre 0.23 et 0.68 et sa teinte se situe entre 0° et 50°.

Dans [17], l'auteur utilise un seuillage neuronal pour la détection de la peau dans des images *HSV*. Les réseaux de neurones sont utilisés aujourd'hui dans plusieurs

domaines tels que la robotique, la classification en biologie, les approximations de fonctions inconnues ainsi que les estimations boursières. Dans le domaine de la vision par ordinateur, ces applications portent principalement sur la compression d'images pour le stockage et la transmission, les informations géométriques environnantes pour les systèmes autonomes et bien sûr la reconnaissance de forme. Un réseau de neurones simple est le perceptron. Chaque perceptron effectue un travail relativement simple : il reçoit des données x pondérées des voisins ou des sources externes et calcule sur cette base un signal de sortie y qui est propagé à d'autres unités comme le montre la figure 1.6.

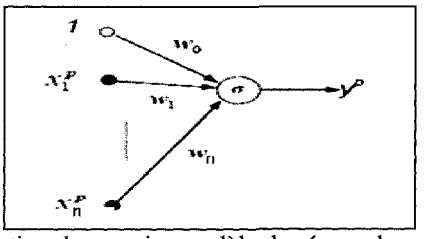

Figure 1.6 Représentation du premier modèle de réseau de neurones (Perceptron).

Toujours dans [17], l'auteur utilise un perceptron multicouche (*MLP*) qui est entrainé avec une banque d'images contenant des peaux extraites de personnes de différentes races. Évidemment, l'auteur utilise des images qui ne représentent aucun sujet humain afin de compléter l'apprentissage de son réseau. Le réseau conçu possède trois entrées, chacune pour un canal de l'espace de couleur *HSV*. Chaque composante du triplet *HSV* est donc fournie en entrée au réseau multicouche et la réponse obtenue en sortie est binaire montrant la présence de peau ou non comme le montre la figure 1.7. Le nombre de neurones de la couche cachée est cinq et il a était déterminé expérimentalement. L'auteur utilise la rétro-propagation de l'erreur comme algorithme d'apprentissage pour trouver les poids optimaux du réseau.

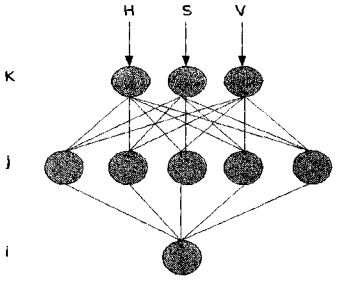

Figure 1.7 Organisation du réseau neuronal présenté dans [17].

1.2.2 Détection de visage et des mains

Certaines méthodes utilisent l'extraction des régions de peau pour la détection des mains et du visage. Ces méthodes nécessitent l'analyse des régions résultantes afin de savoir si elles représentent des parties d'intérêt ou non. Cependant, il existe plusieurs algorithmes qui permettent la localisation du visage et des mains sans étape d'extraction de couleur de peau. Les prochaines sections contiennent quelques-unes de ces méthodes.

1.2.2.1 Les arêtes

Les arêtes sont des points de discontinuités dans la fonction de luminance (intensité) de l'image. Ces informations utiles sont notamment employées pour l'interprétation de scènes et la reconnaissance d'objets. Le principe de base consiste à reconnaître des objets dans une image à partir de modèles de contours connus au préalable. Certains auteurs utilisent la transformée de Hough afin de réaliser cette tâche. Cette dernière permet d'extraire et de localiser des groupes de points respectant certaines caractéristiques. Dans un contexte de détection de visage, ce dernier est représenté par une ellipse dans la carte d'arêtes. Comme dans [18], l'application de la transformée de Hough circulaire va produire une liste de tous les candidats étant des cercles ou des dérivées.

1.2.2.2 L'appariement de gabarit

Cette méthode est certainement une des techniques de détection du visage et des mains la plus simple qui soit. Comme défini dans [19], cette méthode consiste à comparer l'intensité des pixels entre un gabarit prédéfini et plusieurs sous-régions de l'image à analyser. Ce processus correspond en pratique à effectuer plusieurs balayages couvrant toute la superficie de l'image. Les endroits les plus propices à la présence de visages ou des mains seront donc facilement identifiés par des minimums de distance entre le gabarit et l'image sous-jacente. Dans [20], les auteurs détectent des caractéristiques du visage à l'aide de gabarit plus spécialisés (p. ex. : nez, bouche, etc.). L'inconvénient de cette méthode c'est qu'elle implique une recherche intensive dans un vaste espace de solutions possibles (rotation, échelle et translation). Aussi dans [21], Viola et Jones ont créé un

détecteur de visage à temps réel qui traite des images extrêmement rapide avec un grand taux de détection et qui est implanté dans la librairie OpenCV. Ils ont commencé par introduire une nouvelle représentation de l'image nommée «l'image intégrale» qui permet aux caractéristiques utilisées par leur détecteur à être calculées très rapidement. Ils utilisent par la suite un simple et efficace classificateur qui est construit avec l'AdaBoost algorithme d'apprentissage développé en [22] pour sélectionner un petit nombre de caractéristiques visuelles critiques d'un très grand nombre de caractéristiques potentielles. Finalement, ils utilisent une méthode qui combine des filtres dans une "cascade", ce qui permet de rejeter rapidement les régions de l'arrière plan de l'image et utiliser plus de calculs pour le traitement des régions les plus susceptibles de contenir le visage. En résumé, les auteurs élaborent des masques caractéristiques de visages, en travaillant à partir d'une base aussi importante que possible de visages, de façon à disposer d'une représentativité des échantillons. Ils programment par la suite une recherche rapide de zones de l'image dont les contrastes correspondent à ces masques en utilisant une banque de filtres et la réponse à ces filtres.

1.2.2.3 Les réseaux de neurones

Cette méthode présentée précédemment dans le cadre de la détection des régions de peau est aussi utilisée pour la détection des parties du corps telles que le visage et les mains. En effet, [23] présente un réseau de neurones à deux sorties représentant la présence ou l'absence de l'objet recherché dans une sous-région de l'image. Le réseau peut également n'avoir qu'une seule sortie qui se déclenchera lors de la présence d'un visage. Le principe consiste à balayer l'image avec une fenêtre d'attention de dimensions fixes et de réaliser la détection sur les sous-images. Néanmoins, il est encore une fois nécessaire d'effectuer plusieurs balayages à différentes résolutions pour ainsi réaliser une détection suffisamment robuste. Les auteurs utilisent l'ensemble des images de visage et l'ensemble des images de non-visage pour entraîner le réseau de neurones. Le réseau contient autant de cellules d'entrées que de pixels composant la fenêtre de l'image source.

Cette méthode nécessite en effet un très long apprentissage qui dépendra du nombre de neurones sur la couche cachée et de la taille des fenêtres de balayage.

1.2.2.4 Les modèles de Markov cachés

Cette méthode est utilisée depuis plusieurs années dans des domaines autres que la détection et la reconnaissance des parties du corps tels que la bioinformatique, la biométrie et la reconnaissance de la parole et de l'écriture manuscrite. Les chaînes cachées de Markov forment un ensemble de modèles statistiques utilisés pour caractériser les propriétés statistiques d'une image. L'image est divisée en N régions significatives qui sont, par exemple pour la détection du visage, les cheveux, le front, les yeux, le nez et la bouche. Comme présenté dans [20], chacune de ces régions est ensuite assignée à un état S_i dans l'HMM comme le montre la figure 1.8. Pour entrainer l'HMM, chaque échantillon du visage est converti en une séquence de vecteurs d'observation. Les vecteurs d'observation sont construits à partir d'une fenêtre de $W*L$ pixels. En scannant la fenêtre verticalement avec P pixels de chevauchement, une séquence d'observation est construite (voir figure 1.8a). Chaque état est caractérisé par une fonction de probabilité, estimée sur la base des images exemples. Le principe des HMMs, lors de la localisation du visage, est de toujours extraire les mêmes régions de l'image d'entrée et de vérifier si les objets caractéristiques apparaissent dans le même ordre que défini dans le modèle HMM.

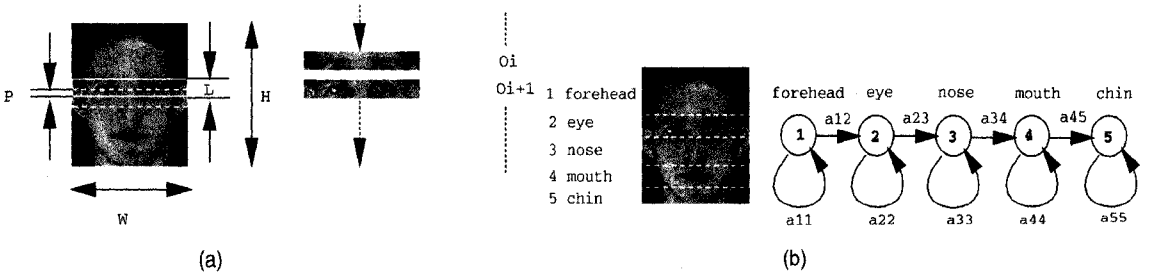

Figure 1.8 La topologie du HMM utilisé pour la détection du visage. a) Vecteurs d'observation, b) Les états cachés de Markov. Figure extraite de [20].

1.2.3 Suivi d'objet

Quelque soit l'objet que l'on veut suivre dans une séquence vidéo, on a besoin d'un modèle pour le décrire : ce modèle peut contenir aussi bien de l'information a priori sur l'objet que de l'information extraite des trames précédentes. Il peut être constitué de descripteurs de couleur, de texture, de forme ou de tout autre type de primitives. En général, il y a deux types d'approches utilisées pour le suivi des objets : les approches par apparences et les approches prédictives. Les prochaines sections décrivent les plus importantes méthodes utilisées dans chacune de ces approches.

1.2.3.1 Approches par apparences

Comme approches d'apparences, on trouve la couleur, la texture et la forme. Ces modèles d'apparences peuvent être utilisés en suivi en comparant les régions candidates ou être utilisés par les principales techniques de suivi telles que le décalage moyen. Chacun de ces modèles sera décrit séparément.

1.2.3.1.1 Couleur

La couleur est typiquement modélisée par un histogramme. Ce dernier calcule le nombre d'occurrences pour chaque couleur (ou classe) dans l'image et peut être construit dans n'importe quel espace de couleurs (*RGB*, *HSV*, *YCbCr*, etc.). Les histogrammes sont relativement invariants selon la translation, la rotation dans l'axe de l'image, le changement d'échelle et les occlusions partielles. Ce qui fait que les histogrammes de couleurs représentent un outil particulièrement intéressant pour la reconnaissance d'objets ayant une position et une rotation inconnues par rapport à la scène. L'utilisation des histogrammes de couleurs pour le suivi des objets exige une étape de comparaison. En effet, pour savoir si deux images ou deux régions ont la même distribution de couleurs, il suffit de comparer leurs histogrammes de couleurs. Il existe différentes mesures pour comparer des histogrammes. Dans [24], les auteurs définissent les notions d'intersection et de distance entre deux histogrammes. L'intersection d'histogrammes vérifie la qualité de l'inclusion de l'histogramme modèle $h(M)$ dans l'histogramme image $h(I)$. Il y a correspondance si tous (ou presque) les pixels des K classes de l'histogramme $h(M)$ sont

inclus dans les K classes de l'histogramme $h(i)$. L'intersection et la correspondance sont calculées selon les équations

$$\begin{aligned} \text{intersection}(h(I),h(M)) &= \sum_{j=1}^{K} \min\{h(I)[j],h(M)[j]\}, \\ \text{correspondance}(h(I),h(M)) &= \frac{\sum_{j=1}^{K} \min\{h(I)[j],h(M)[j]\}}{\sum_{j=1}^{K} h(M)[j]}. \end{aligned} \tag{1.11}$$

Pour calculer la distance entre deux histogrammes et en mesurer la similarité, plusieurs mesures de distances peuvent être utilisées. Parmi ces distances, on trouve la distance pâté de maison '*City block*' (D_1) et euclidienne (D_2) présentées dans l'équation 1.12. Ces distances considèrent l'histogramme comme un vecteur.

$$\begin{aligned} D_1(h(i),h(M)) &= \sum_{j=1}^{K} |h(i)[j]-h(M)[j]|, \\ D_2(h(i),h(M)) &= \sqrt{\sum_{j=1}^{K} (h(i)[j]-h(M)[j])^2}. \end{aligned} \tag{1.12}$$

La distance MDPA est aussi utilisée pour la comparaison des histogrammes de couleurs. Elle considère un histogramme comme un vecteur et permet de savoir le prix minimum pour passer d'une distribution à une autre. Cette distance a été utilisée dans le cadre de notre recherche et elle est expliquée dans le chapitre 2.

D'autres distances considèrent l'histogramme comme une distribution discrète de probabilités comme la distance de Bhattacharrya qui est calculée selon l'équation 1.13.

$$D(h(i),h(M)) = -\log \sum_{j=1}^{K} \sqrt{P(h(i)[j])P(h(M)[j])}. \tag{1.13}$$

La couleur peut aussi être modélisée par une signature. Une signature est une représentation compressée et sans perte d'un histogramme [25]. On construit une signature en éliminant les classes nulles d'un histogramme. On enregistre par la suite le nombre w_j de pixels qui appartiennent à la classe j de l'histogramme et on associe à chaque classe j un poids m_j qui correspond à la moyenne des valeurs des pixels dans la

classe. Un exemple de construction d'une signature à partir d'un histogramme de couleur est présenté dans la figure 1.9.

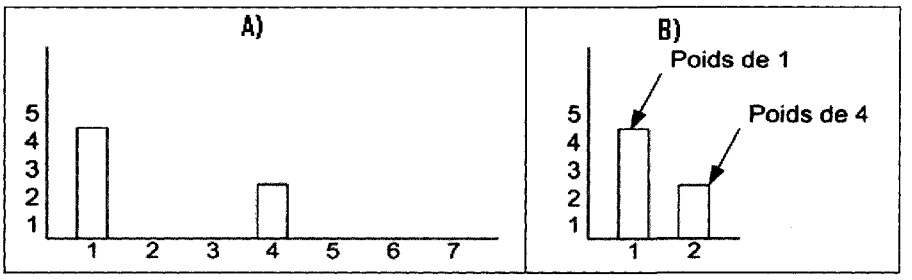

Figure 1.9 Construction d'une signature à partir d'un histogramme. A) Un exemple d'histogramme de couleur, B) La signature équivalente à l'histogramme présenté en A).

La modélisation par histogramme ou signature n'est pas précise, car la position des couleurs n'est pas prise en compte. Pour cela, d'autre modèle d'apparence sont utilisés. La modélisation par texture s'avère plus précise, car elle prend en compte la position relative des couleurs directement ou indirectement et sera présentée dans la section qui suie.

1.2.3.1.2 Texture

Il existe différente méthode pour étudier la texture. Dans [26], les auteurs définissent la densité des arrêtes qui consiste à mesurer le nombre d'arêtes par unité de surface. Les arêtes sont caractérisées par un changement rapide d'intensité. Cette caractéristique de texture a été utilisée dans le cadre de notre étude afin de nous permettre la localisation de la main dans le bras et elle est expliquée plus en détail au chapitre 2.

Une autre technique pour modéliser la texture est la banque de filtres présentée dans [27]. Cette technique consiste à faire la convolution de l'image avec différent filtres, chacun permettant d'extraire une propriété différente. Si une région d'image a une distribution spatiale semblable au filtre, la réponse de la convolution avec le filtre sera grande.

Une autre méthode pour décrire la texture est la matrice de co-occurrences introduite par [28]. Cette dernière est une matrice qui enregistre le nombre de fois que

deux couleurs (valeurs d'intensité) similaires ont la même position relative dans l'image. Ce qui nous permet d'avoir une information sur la distribution spatiale des couleurs. L'équation 1.14 décrit une matrice de co-occurrence *C(i,j)* définie pour la relation spatiale *(dx,dy)*.

$$C(i,j) = \left|\{(x,y) \in R | I(x,y) = i \wedge I(x+dx, y+dy) = j\}\right| \qquad (1.14)$$

La figure 1.10 représente un exemple de construction d'une matrice de co-occurrence pour la relation spatiale (dx=0, dy=1).

Figure 1.10 Matrice de co-occurrences construite à partir d'image exemple. A) Matrice de co-occurrences, B) Image exemple.

Pour la comparaison des matrices de co-occurrences, les mêmes techniques que pour les histogrammes adaptées pour deux dimensions sont utilisées. Un autre modèle pour analyser l'apparence d'un objet est la forme. Ce dernier est décrit dans la section suivante.

1.2.3.1.3 Forme

Une manière de suivre un objet est d'analyser sa forme. La modélisation par moments qui donne la distribution des points composants l'objet est l'une des méthodes les plus utilisées pour décrire des formes. Dans [29], les auteurs présentent les moments de Hu comme étant un descripteur de forme invariant vis-à-vis les translations, la rotation et le changement d'échelle. Ces moments sont utilisés dans le cadre de notre travail pour le suivi de la tête. L'idée de l'utilisation d'un descripteur de forme pour le suivi du visage

est due au fait que la forme du visage et différente de celle des mains et varie moins lors d'une activité humaine. Ce descripteur de forme est expliqué plus en détail au chapitre 2.

1.2.3.1.4 Le décalage vers la moyenne (Mean shift)

Dans [30], cette méthode de suivi utilisant les modèles d'apparence (couleur, texture et forme) est définie comme étant une méthode efficace et non paramétrique pour chercher des objets basée sur l'estimation de la densité du noyau. Soit les données un ensemble fini A inclus dans l'espace euclidien X. La moyenne d'échantillons à $x \in$ X est présentée dans l'équation

$$sm(x) = \frac{\sum_a K(a-x)w(a)a}{\sum_a |K(a-x)w(a)|}, a \in A, \tag{1.15}$$

avec K est une fonction de noyau, et w est un poids qui peut être négatif. La différence sm(x)-x s'appelle le vecteur moyen de décalage. Le mouvement répété des points de repères aux moyennes d'échantillon s'appelle l'algorithme de décalage vers la moyenne.
Un noyau couramment utilisé est le noyau gaussien dont le profil est défini par l'équation 1.16.

$$k(x) = \frac{1}{2\pi}\exp(-\frac{1}{2}\|x\|^2) \tag{1.16}$$

Dans [31], les auteurs emploient le coefficient de Bhattacharyya pour mesurer la similitude de deux noyaux d'histogrammes de couleurs représentant l'image d'objet et l'image de candidat respectivement. L'information sur l'objet suivi est mise à jour pour chaque changement majeur qui s'effectue au cours du temps.

1.2.3.2 Approche prédictives

Plusieurs algorithmes de suivi se basent sur des probabilités et des hypothèses. Ils développent un modèle de prédiction qui va permettre le suivi de l'objet d'intérêt. Parmi ces méthodes, on trouve le filtre de Kalman et les filtres particulaires. Chacune de ces méthodes sera décrite séparément.

1.2.3.2.1 Filtre de Kalman

Le filtre de Kalman est un estimateur récursif qui se base sur seulement l'état précédent et les mesures actuelles pour estimer l'état courant et ne requiert pas l'historique des observations et des estimations. Le filtre de Kalman a deux phases distinctes : la phase de prédiction et celle de la mise à jour. La phase de prédiction utilise l'état estimé de l'instant précédent pour produire une estimation de l'état courant. Dans l'étape de mise à jour, les observations de l'instant courant sont utilisées pour corriger l'état prédit dans le but d'obtenir une estimation plus précise. En suivant un objet, les informations telles que la vitesse et la position sont filtrées du bruit qu'elles peuvent contenir pour ainsi permettre une meilleure prédiction de la future position de l'objet en mouvement. Outre le filtre de Kalman pour suivre les objets, plusieurs auteurs utilisent le filtrage particulaire. Ce dernier est présenté dans la section qui suit.

1.2.3.2.2 Filtre de particules

Comme défini dans [32], les techniques de filtrage particulaire sont des méthodes utilisées pour l'estimation récursive du vecteur d'état d'un système stochastique Markovien. En effet, on se place dans le cadre Bayésien pour suivre des objets lorsque les densités de probabilité a posteriori $P(X_t \mid Z_t)$ et le modèle d'observation $P(Z_t \mid X_t)$ ne sont pas nécessairement gaussiennes. L'objet suivi est caractérisé par son vecteur d'état X_t, et les observations du temps t = 0 jusqu'au temps t sont définies par le vecteur Z_t. L'idée du filtrage particulaire est d'approcher la distribution de probabilité de l'état de l'objet par un ensemble d'échantillons associés à des poids. Chaque échantillon, aussi appelé particule, est un élément qui représente un état hypothétique de l'objet *s*, auquel on associe un poids π, représentant sa vraisemblance par rapport au modèle. On peut écrire l'ensemble de la façon suivante :

$$S = \left\{ \left(s^{(i)}, \pi^{(i)}\right),\ i = 1,...,n \right\} \text{ où } \sum_{i=1}^{n} \pi^{(i)} = 1. \tag{1.17}$$

L'évolution de l'ensemble est obtenue en propageant chacun des échantillons selon un modèle de mouvement. On attribue ensuite un poids à chaque particule en fonction des observations, et on détermine l'état moyen du système à chaque étape par :

$$E[S] = \sum_{i=1}^{n} \pi^{(i)} s^{(i)}. \tag{1.18}$$

Un des avantages du filtrage particulaire est que l'on modélise l'incertitude : en sortie de l'algorithme, il n'y a pas une solution unique, mais un ensemble de particules représentant la densité de probabilité du vecteur d'état dans l'image. Par conséquent, ce type de filtrage peut constituer une approche robuste en cas de bruit ou d'occultation.

La limite principale de ces approches prédictives est qu'on suppose des mouvements facilement prédictibles. Ces méthodes sont moins fiables lorsque des changements brusques de direction des objets suivis se produisent. Dans ces cas, les prédictions sont souvent fausses. Pour augmenter la robustesse de ces modèles de suivi et d'interprétation des mouvements, plusieurs auteurs les ont combinés pour avoir une meilleure efficacité de suivi. En effet, dans [30] le filtre de Kalman a été combiné au Mean-shift pour permettre le suivi des blobs d'intérêts. Dans [33], les auteurs combinent la méthode Mean-shift avec un filtre particulaire pour lui ajouter une composante prédictive et augmenter l'efficacité du suivi dans leur système. D'autres auteurs ont combiné les approches prédictives à celles par apparence pour pouvoir suivre des objets dans des scènes. En effet, dans [34], le suivi adaptatif des objets non rigides s'effectue en combinant les histogrammes de couleurs au filtrage particulaire. Dans [35], les auteurs proposent un nouveau modèle de couleur combiné à l'algorithme Mean-shift pour un suivi robuste et en temps réel des objets.

Pour ce qui est du suivi, les approches prédictives ont montré des limites surtout dans le cas où les mouvements des objets d'intérêts sont difficiles à prévoir. C'est le cas des mains dans le cadre d'une activité telle que la prise de médicaments. Pour cette raison, nous optons pour l'utilisation de méthodes par apparence. En effet, on utilise dans un premier temps la couleur pour détecter les régions de peau contenues dans chacune

des trames. Par la suite un descripteur de forme est utilisé pour la localisation et le suivi du visage. Le suivi des mains s'effectue en exploitant les propriétés de contours et du centroïde des régions. La détection et le suivi des bouteilles de médicaments sont effectués en combinant des techniques basées sur la couleur et la forme. Toutes ces techniques sont expliquées en détails dans le chapitre 2 de ce mémoire.

1.3 Reconnaissance d'activité humaine

Cette section présente les méthodes élaborées dans les recherches antérieures afin de reconnaitre différentes activités humaines. Dans [36], les auteurs ont développé un modèle caché de Markov pour identifier les activités dans une salle à manger. La topologie du modèle est déterminée en estimant combien de différents états sont impliqués dans des activités dans une salle à manger. Ils utilisent quatre états pour décrire les étapes de l'activité. Deux états modélisent les deux mouvements qui peuvent être employés pour marquer le début et la fin du repas. En effet, l'état q_1 représente le déplacement des mains vers la tête, tandis que l'état q_2 modélise le mouvement relatif dans la direction opposée. Un troisième état est créé pour modéliser les mouvements qui sont non-reliés à l'activité de manger. Pour les cas où aucun mouvement n'est détecté, un état dc « don't care» est aussi utilisé. Donc, les activités de prise de repas sont apprises par l'HMM montré à la figure 1.11.

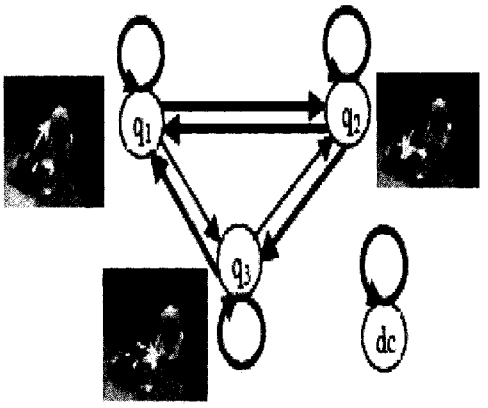

Figure 1.11 Topologie du HMM utilisé pour l'analyse de l'activité de prise de repas. Figure extraite de [36].

Dans notre cas, le nombre de scénarios pour la prise de médicaments est plus grand et par conséquent le nombre d'états sera largement supérieur à celui utilisé dans cette étude ce qui nécessitera une longue phase d'apprentissage pour l'entrainement et la construction d'un HMM pour détecter l'activité de prise de médicaments.

Dans [37], les auteurs ont développé un algorithme en temps réel qui permet le suivi des objets multiples dans des environnements complexes pour pouvoir identifier des activités. L'identification d'activités est basée sur la trajectoire et le comportement (séparation et fusion) des régions (blobs) à travers les séquences capturées. L'algorithme débute par une extraction d'arrière-plan et les pixels de premier plan sont détectés en se basant sur le contraste de luminance. Ce dernier représente la différence relative entre la luminance de l'objet et la luminance de l'arrière-plan. Dans leur approche, les auteurs utilisent l'espace de couleur YUV qui est obtenu selon les équations

$$\begin{aligned} Y &= 0.299*R + 0.587*G + B, \\ U &= 0.436*(B-Y) / (1-0.114), \\ V &= 0.615*(R-Y) / (1-0.299). \end{aligned} \tag{1.19}$$

Les pixels d'avant-plan sont regroupés en régions (blobs). Pour chaque blob détecté, la trajectoire est calculée à l'aide des intersections des rectangles englobant des blobs de deux images consécutives. La trajectoire tient aussi en compte les séparations/fusions de blobs aussi établis à l'aide des rectangles englobant et d'une matrice d'intersection de blobs. Connaissant le comportement et les trajectoires des blobs, les auteurs utilisent des règles pour pouvoir détecter certains événements. Par exemple, pour la détection de bagages abandonnés il faut détecter qu'un blob se sépare en deux blobs. Par la suite, un des deux blobs reste stationnaire sur une période de temps définie et l'autre s'éloigne. Un problème avec cet algorithme c'est qu'il nécessite une extraction d'arrière-plan parfaite (ou presque) pour la détection des blobs. D'autre part, l'algorithme peut être utilisé pour la détection de scénarios simples et il ne peut s'appliquer à des scénarios complexes.

Dans [38], les auteurs ont proposé une nouvelle approche pour l'interprétation des séquences vidéo basée sur les modèles déclaratifs des activités. Ils présentent une

approche à deux niveaux afin de comprendre ce qui se produit dans la scène. Un niveau «événement» représentant les changements significatifs dans l'état de la scène et un niveau «scénario» dont l'objectif est de reconnaître des situations modélisées par des combinaisons d'événements. Les auteurs introduisent la notion de *fait*. Ce dernier correspond à un objet défini par une série d'attributs tels que le nom, la date, le type, etc. la hiérarchie des faits est présentée dans la figure 1.12. Chaque fait abstrait est modélisé par un ensemble d'attributs correspondant aux faits concrets impliqués et aux conditions permettant la production du fait. Un scénario est donc reconnu si toutes les conditions de son modèle sont respectées.

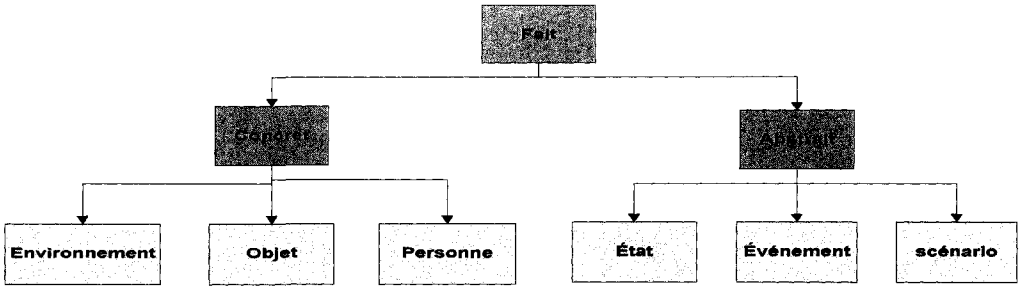

Figure 1.12 Hiérarchie des *faits*.

L'algorithme a été utilisé pour la détection d'événements simples tels que le rapprochement et l'éloignement d'une personne d'un guichet pour la reconnaissance d'un scénario de vandalisme dans les guichets de billets dans une station métro. En effet, l'algorithme semble bien fonctionner pour la détection de scénarios relativement simples. Cependant, il est difficile de l'utiliser avec des scénarios plus complexes.

Dans [39], les auteurs ont développé un système de représentation des événements dans une séquence vidéo en utilisant des réseaux de Petri et ont par la même occasion évoqué les avantages de l'utilisation de ces derniers pour la représentation et la reconnaissance des événements. Parmi ces avantages, on trouve la représentation des événements de façon séquentielle, simultanée et synchronisée. Par conséquent, pour la détection de l'activité humaine qui est dans le cadre de notre étude la prise de médicaments, on a utilisé un réseau de Petri. Le principe de ce dernier est bien détaillé dans le chapitre 2.

CHAPITRE 2 MÉTHODOLOGIE

Dans ce chapitre, on décrit les méthodes utilisées afin d'atteindre les objectifs du projet. On propose un aperçu de l'approche à la section 2.1. Les hypothèses, les contraintes et le cadre d'application de notre système seront présentés à la section 2.2. La méthode utilisée pour la détection des régions de la peau contenues dans chacune des trames de la séquence vidéo est décrite à la section 2.3. La section 2.4 présente la technique adoptée pour la détection et la gestion des occlusions entre les parties du corps suivies. Les sections 2.5 et 2.6 révèlent les méthodes utilisées pour la détection et le suivi du visage et des mains. À la section 2.7, on explique comment les bouteilles de médicaments sont localisées et identifiées dans chacune des trames. On termine ce chapitre par la présentation du réseau de Petri construit pour la détection de l'activité humaine qui est dans notre cas la prise de médicaments à la section 2.8.

2.1 Aperçu de la méthode

La figure 2.1 présente le pseudo-code schématique global du programme développé.

```
En entrée, on reçoit le nom de la séqeunce vidéo qu'on veut analyser.
Initialisation du réseau de Pétri
POUR chacune des trames de la séqeunce vidéo d'entrée FAIRE
    Sauvegarde de l'image couleur
    Transformation de l'espace de couleur RGB à HSV
    Seuillage et segmentation pour la détection des régions de peau humaine
    Vérification des occlusions entre les parties du corps qui nous intéressent (Visage+Mains)
    SI le visage n'est pas en occlusion ALORS
        SI Trame initiale de la séquence ALORS
            Hypothèse pour la localisation du visage
        SINON
            Détection et suivi par moment de Hu d'ordre 2
    SI les bras ne sont pas en occlusion ALORS
        Détection des bras contenus dans la trame
        Détection des contours des bras
        Localisation des mains dans les bras
    Détection et identification des bouteilles de médicaments
    Vérification des occlusions entre les bouteilles et les mains
    Mise à jour du réseau de Pétri et vérification de la prise de médicaments
```

Figure 2.1 Pseudo-code schématique de l'algorithme

En entrée, le système reçoit le nom de la séquence vidéo à analyser. Pour chacune des images (trames) de la séquence, l'espace de couleur *HSV* avec des seuils prédéfinis est

utilisé pour détecter les régions de peau. Le nombre de régions de peau extraites dans chacune des trames nous permet de détecter les occlusions présentes dans la séquence et de mettre à jour l'état du système. Par la suite un descripteur de forme est utilisé pour le suivi du visage. La localisation de ce dernier au début de la séquence se fait en utilisant quelques hypothèses. Le suivi des mains s'effectue en exploitant les propriétés de contours et du centroïde des régions. L'identification des bouteilles de médicament se fait en combinant les histogrammes de couleurs et un descripteur de forme et le suivi de ces derniers se base sur la propriété du centroïde. Nos algorithmes de localisation et de suivi des parties du corps et des bouteilles de médicaments sont appliqués pour la détection de l'activité humaine. Dans notre cas, on a utilisé un réseau de Petri afin de reconnaître la prise de médicament. La transition des jetons d'une place à l'autre ne se produit que si les événements durent un certain nombre de trames.

2.2 Hypothèses et cadre d'application

Les médicaments peuvent être placés dans différents contenants et être pris dans plusieurs places et par conséquent les façons de les prendre sont multiples et un système de détection de prise de médicaments, indépendant d'un contexte précis, est très complexe. Afin de simplifier le problème, on considère que les médicaments se trouvent dans des bouteilles et que la personne les prend toujours au même endroit, soit à la table et face à la caméra. Ceci permettait d'avoir une vue des bouteilles de médicaments et des parties du corps de la personne tout au long de l'activité. On suppose également qu'une seule personne assise sur une chaise prend ses médicaments devant la caméra pour pouvoir appliquer nos algorithmes de détection et de suivi du visage et des mains et que cette personne est en face de la caméra à la trame initiale de la séquence pour localiser correctement le visage. Aussi, on suppose que les mains et le visage sont visibles et ne sont pas en occultation ou en contact dans la trame initiale présentée au système.

Pour la localisation et le suivi des parties du corps, on s'est basé sur les caractéristiques de la couleur de la peau afin de détecter les régions de peau contenues dans chacune des trames de notre séquence d'entrée. Pour pallier aux inconvénients de cette méthode, on a évité que la couleur de l'arrière plan, des vêtements des personnes et

des objets présents dans la séquence (incluant les bouteilles de médicaments) soient semblables à la couleur de la peau lors de la prise des séquences vidéo. Un exemple de séquence avec des objets (tables de bureau) qui sont classifiés comme étant de la peau est présenté à la figure 2.2. Les objets de couleur semblable à celle de la peau sont par conséquent enlevés de l'environnement afin de réduire le nombre de régions susceptibles de contenir des parties du corps.

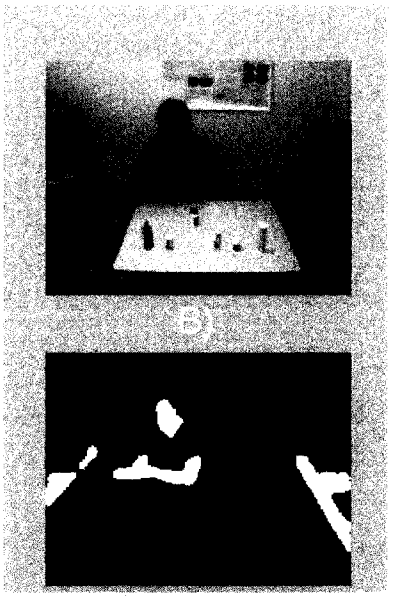

Figure 2.2 Exemple de fausse classification. A) trame source, B) Détection des régions de peau contenues dans cette trame.

Pour l'emplacement de la caméra, plusieurs positionnements ont été envisagés. On a commencé par placer la caméra directement en face de la personne comme illustré à la figure 2.3A. Cet emplacement a comme avantage d'avoir une vue de face du visage mais augmente les occlusions entre les mains et les bouteilles de médicaments. En effet, une main se trouvant derrière une bouteille de médicaments sera perçu comment étant en contact avec la bouteille. Aussi selon notre algorithme de détection des régions de la peau, la partie occultée de la main ne sera pas détectée (figure 2.3B) et par conséquent les contours de cette dernière ne seront pas correctement définis. Placer la caméra du côté de la personne compliquera le suivi des mains puisqu'à ce point de vue de la caméra, les mains seront toujours occultées. Aussi, un emplacement de la caméra en haut de la

personne ne peut pas être adopté puisque le visage ne sera pas visible et par conséquent le suivi de ce dernier est impossible. Finalement, on a placé la caméra devant la personne et légèrement au-dessus de sa tête, comme illustré à la figure 2.3C, pour diminuer les occlusions entre les objets d'intérêts qu'on veut détecter et suivre. Concernant la distance entre la caméra et la personne, les résultats sont meilleurs lorsque la caméra est proche de la personne. Cette distance peut variée entre deux et cinq mètres. Dans ce cas, la taille des objets est plus grande, ce qui facilite leur localisation et leur suivi.

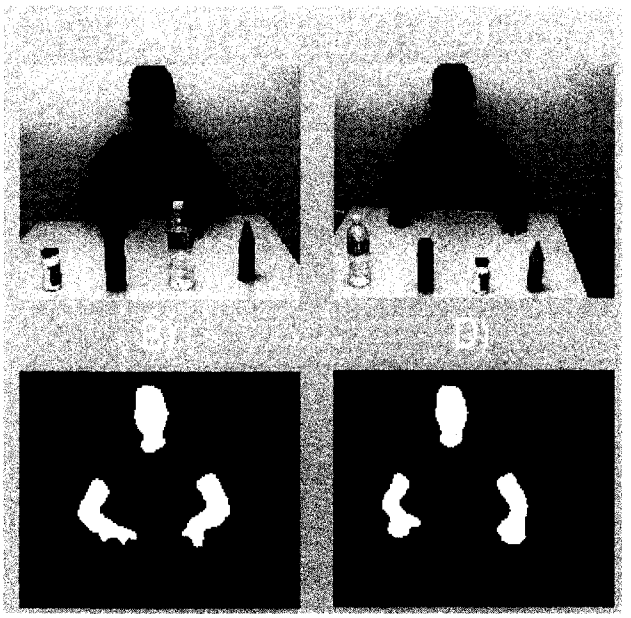

Figure 2.3 Vues obtenues selon la position de la caméra A), de face C), de face surélevée B), D), Détection des régions de la peau contenues dans les images A et C respectivement.

Dans le cas où des objets autres que les bouteilles de médicaments sont posés sur la table (comme illustré à la figure 2.3), notre algorithme les associe aux bouteilles de médicaments ayant la forme et la couleur la plus proche de ces derniers. Cependant, ces objets sont enlevés de la table lors des séquences prises et des tests effectués sur ces séquences. Finalement, notre système ne fonctionnera pas si la personne ne porte pas de chandail et par conséquent, on suppose que la personne porte au moins un débardeur pour pouvoir applique nos algorithmes de détection et de suivi des parties du corps.

2.3 Détection des régions de peau

Dans notre étude, pour pouvoir localiser et suivre le visage et les mains, on s'est concentré sur la méthode basée sur la couleur de la peau étant donné que celle-ci possède des avantages importants tels que la rapidité et l'efficacité de la détection des régions de la peau [19]. Avec cette méthode et après seuillage et segmentation des régions de la peau, on pourra détecter le visage et les mains contenus dans l'image source. Cette étape est très importante puisqu'une mauvaise localisation du visage et des mains va diminuer la possibilité de reconnaître l'activité humaine.

Comme mentionné au chapitre 1, la peau humaine est souvent représentée par une portion d'un espace de couleur particulier et il est par conséquent possible d'extraire les pixels dont la couleur peut s'apparenter à celle de la peau. On commence donc par obtenir une image RGB de notre séquence vidéo d'entrée. Un exemple d'image obtenu est présenté à la figure 2.4.

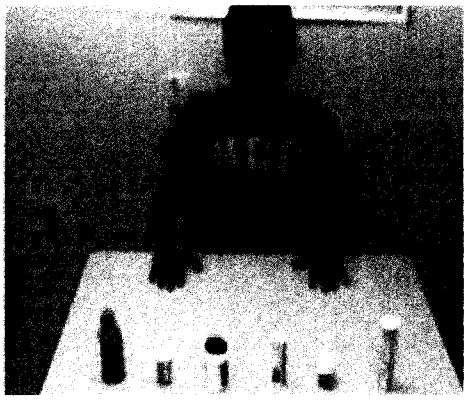

Figure 2.4 Image représentant une trame de la séquence vidéo test.

Par la suite, on transforme notre image à un modèle de couleur plus approprié pour la détection de la peau. Dans le cadre de notre recherche, on a utilisé l'espace de couleur *HSV (H : teinte, S : satuation, V : luminance).* L'avantage de ce dernier pour la détection des couleurs réside dans le fait que le canal *V* représente la luminance permettant ainsi de séparer la chrominance de la luminosité. Ainsi, la teinte, la saturation et la valeur deviennent un espace de couleur alternatif et n'importe quelle couleur peut être décomposée en ces trois composantes. Cette transformation est illustrée par la figure 2.5.

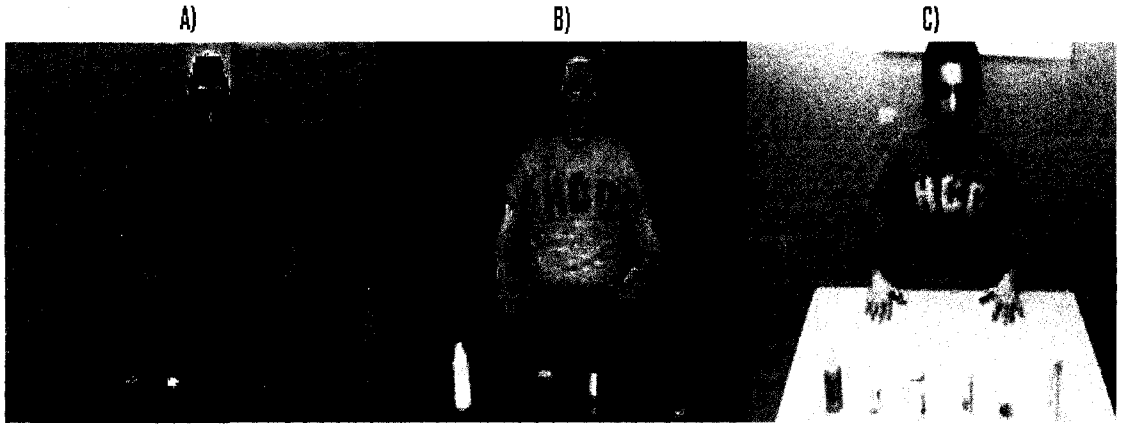

Figure 2.5 Transformation de l'espace de couleur RGB à l'espace de couleur HSV. A) H, B) S, C) V

Certains auteurs ont proposées des seuils appropriés pour ce type d'opération. Lors de notre travail, on a utilisé pour les canaux H et S, les seuils utilisés dans [19]. En effet, dans ce dernier, les auteurs ont ignoré le canal V lors du seuillage. Les seuils utilisés dans notre recherche pour la détection des pixels de la peau sont illustrés au tableau 2.1.

Tableau 2.1 Seuils utilisés pour l'espace de couleur *HSV* afin de détecter les pixels de la peau

Canal	Seuil inférieur	Seuil supérieur
Hue ou Couleur pure (H)	0 (0 °)	0.1 (36°)
Saturation (S)	0.2	1
Valeur (V)	0.2	0.8

Pour qu'un pixel soit étiqueté comme étant de la peau, il doit donc se repérer à l'intérieur de tous les intervalles précités. Le résultat de cette classification est présenté dans la figure 2.6.

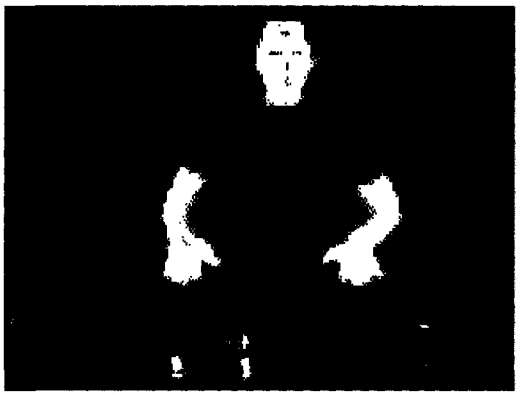

Figure 2.6 Détection des pixels de la peau de la trame avec le modèle HSV

Une fois la classification des pixels de la peau effectuée, on applique un filtre pour enlever les pixels dispersés représentant les fausses classifications. En effet, afin de filtrer l'image obtenue, on a utilisé premièrement la fonction *fspecial* de Matlab avec le type *Disk* de rayon 5 pour appliquer par la suite la fonction *imfilter*. Ces fonctions permettent d'effectuer un lissage sur l'image en noir et blanc en prenant tous les pixels dans le disque de rayon 5 et en leur affectant la moyenne de ces points au pixel central. Le résultat du lissage est une image en tons de gris. Par la suite, on a segmenté la nouvelle image par composantes connectées qui consiste à décomposer l'image en des régions de 0 et 1 avant de numéroter de 1 à n les régions de pixels de valeur 1. Dans un premier temps, on a utilisé la fonction *im2bw* en conjonction avec *graythresh* pour convertir notre image filtrée en une image noir et blanc. Le niveau de gris servant de seuil à cette fonction est déterminé par la fonction *graythresh* qui réalise le calcul de la meilleure valeur de seuil selon le critère d'Otsu. Ce dernier examine la distribution des niveaux de gris pour fixer automatiquement un seuil qui permettrait de classifier les pixels. L'idée de la méthode d'Otsu développée dans [40] est de minimiser la variance intra-groupe pour trouver un seuil maximisant la séparabilité entre les classes. Dans un deuxième temps, la fonction *bwlabel* permettant d'effectuer la segmentation par composantes connectées a été utilisée en conjonction avec les fonctions *regionprops* et *ismember*. En effet, après l'isolation des différents groupes de pixels connectés représentant la peau, communément appelés blobs ou nuées de pixel, on a utilisé la

fonction *regionprops* afin d'extraire les propriétés de chacun des blobs. La propriété de l'aire a été utilisée par la suite en conjonction avec la fonction *ismember* pour ne laisser que les trois blobs ayant les plus grandes aires pour la première trame et rejeter les blobs qui possèdent une aire trop faible pour les suivantes. Le seuil utilisé pour les trames suivantes est calculé à partir de la moyenne entre l'aire la plus petite des régions de la peau et l'aire la plus grande des régions qui ne représentent pas la peau dans la première trame. La figure 2.7 illustre le résultat du seuillage et de la segmentation des régions de peau avec les techniques mentionnées précédemment pour des trames de séquences vidéo différentes.

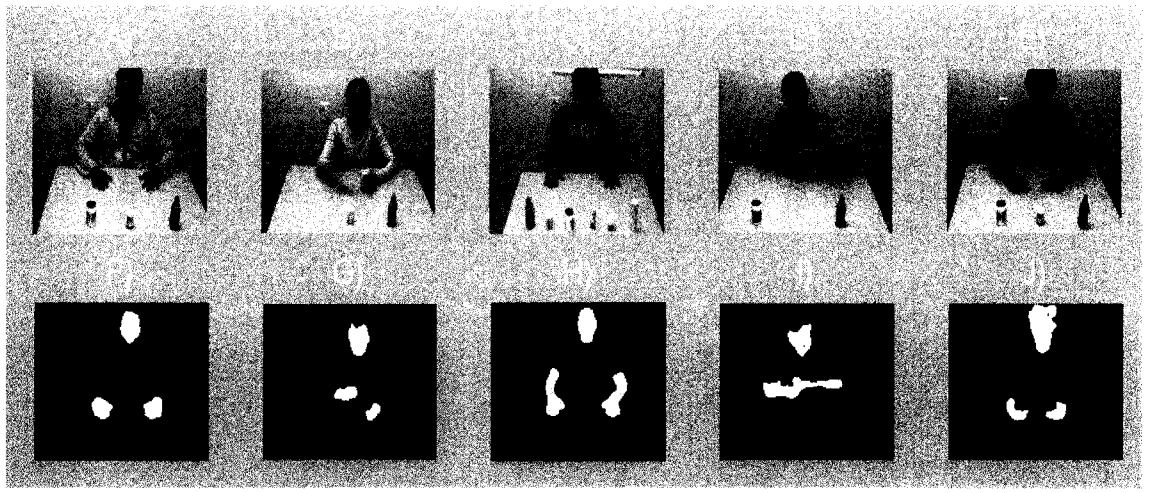

Figure 2.7 Exemple d'extraction des régions de la peau. A), B), C), D), E), Trames des séquences vidéos originales, F), G), H), I), J), Détection des régions de la peau contenues dans chacune des trames.

Il est intéressant de remarquer les fausses détections pour les cheveux (figure 2.7J) ainsi que les régions non détectées (p. ex : petite partie du front (figure 2.7G)). On remarque que dans la séquence présentée à la figure 2.7E, les cheveux ont une couleur semblable à celle de la peau et par conséquent ils seront classifiés comme région de peau. Dans le cas des régions non détectées, on peut voir que certaines zones reflètent davantage la lumière et semblent ainsi plus éclairées. Ce changement de l'intensité de la lumière modifie la couleur de la région dans l'espace HSV. Dans [19], les auteurs on montré que lorsque la saturation est basse et que la luminance est élevée ou l'inverse, la couleur tend vers le blanc, échappant donc aux seuils de détection. Pour limiter le nombre de pixels qui ne

seront pas détecté à cause du reflet, on a rétrécit l'intervalle de classification de la valeur comme le montre le tableau précédent. Dans le cas général, on peut dire que les seuils qu'on a choisi ne provoquent aucun conflit avec des couleurs différentes que celles de la peau, peuvent détecter différents types de peau dans différentes conditions d'illumination et permettent de localiser le visage malgré le port de lunette ou casquette (figure 2.7D). Une fois les parties du corps détectées, notre système procède à la vérification des occlusions. Cette partie est détaillée dans la section suivante.

2.4 Occlusion entre les régions de la peau

Lorsqu'on suit des objets mobile, il est indispensable de détecter les occlusions de ces derniers afin d'analyser leurs comportements et comprendre ce qui se passe dans la séquence. Dans le cadre de notre travail, lorsqu'on parle d'occlusion cela comporte les occultations et les collisions. Une occultation survient lorsqu'un des objets suivi occulte partiellement ou totalement un autre objet suivi sur l'image. Cependant, lorsque les régions suivies se fusionnent dans une image, cela ne traduit pas forcément une occultation mais peut traduire une collision (un contact entre les régions). Généralement, pour différencier une collision d'une occultation, la position 3D est utilisée. Dans notre recherche, le fait de ne pas utiliser un positionnement en trois dimensions nous empêche de distinguer les deux.

Comme mentionné dans la section 2.2, on suppose que les parties du corps (mains + visage) sont visibles et ne se trouvent pas en occlusion dans la trame initiale présentée au système. La segmentation fait en sorte de ne laisser que les trois blobs représentants ces parties dans la première trame puisque tous les objets de couleurs semblables à celles de la peau sont enlevés de la scène. Par la suite, on analyse le nombre de régions de peau extraites dans chacune des trames afin de gérer les occlusions présentes dans la séquence d'entrée. Ceci revient à comparer le nombre de régions de peau détectées dans la trame courante F_t avec celui de la trame précédente F_{t-1}. Cette comparaison nous permet de repérer la présence des occlusions ainsi que le nombre de régions occultées comme le montre la figure 2.8.

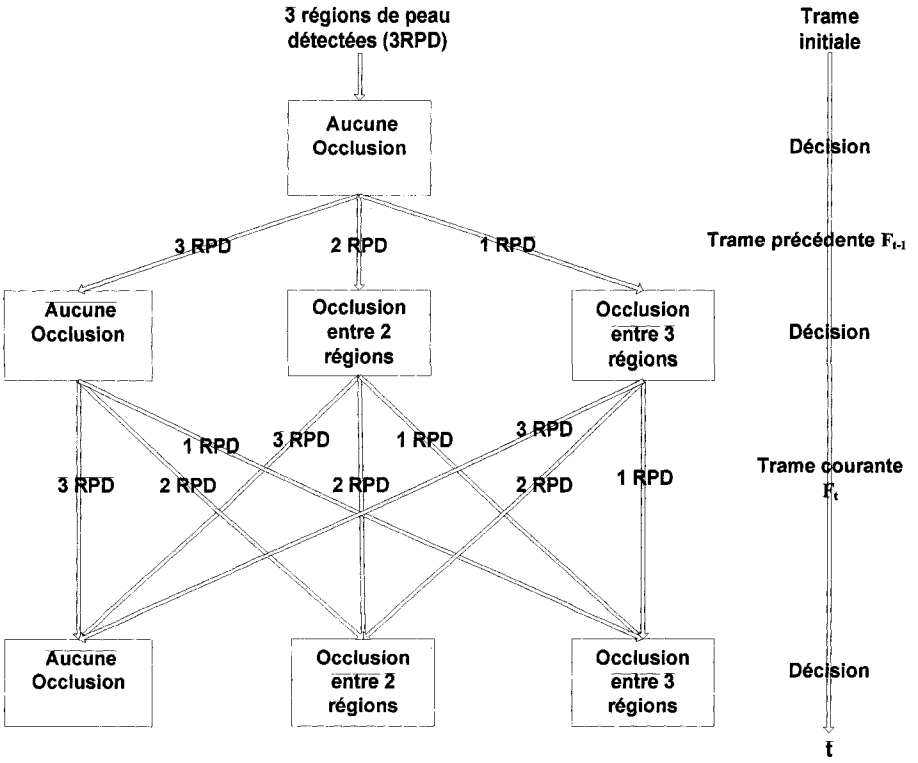

Figure 2.8 Gestion des occlusions en se servant des régions de peau extraites.

Une fois une occlusion détectée, on compare les distances entre les centroïdes des régions de peau de la trame précédente pour savoir les deux régions les plus proches dans F_{t-1} et qui sont maintenant en occlusion dans F_t. Les techniques de localisation et de suivi des régions de peau seront présentées dans les sections qui suivent. Le problème des occlusions est un des problèmes les plus répandus et les plus difficiles à traiter lors du suivi d'objets multiples. Dans [41], les auteurs définissent deux approches pour le suivi des objets en occlusions. L'approche Merge-split (MS) qui détecte les événements de fusion et de séparation des blobs et qui ne fait rien durant l'occlusion et l'approche Straight-through (ST) qui ne détecte pas ces événements et dans laquelle le système continue de suivre tous les objets, même en occlusion. Dans [42], les auteurs énoncent les différents problèmes lors du traitement des occlusions. Parmi ces problèmes, on trouve la détection de l'événement d'occlusion ainsi que le suivi correct des objets en occlusion et l'objet occultant durant l'occlusion. Pour pallier ce dernier problème, il faut disposer d'un

indice discriminant entre les objets suivis. La plupart des auteurs utilisent la couleur comme indice. Dans notre cas, les objets suivis sont des parties du corps et ne possèdent pas de couleur discriminante. Cependant et afin de limiter les fausses détections et erreurs de suivis des différentes régions de peau lors des occlusions, on a adopté l'approche MS. Le suivi des régions d'intérêt est donc suspendu pendant la durée de l'occlusion. Lorsque les objets se séparent, l'algorithme de détection et suivi de ces objets se poursuit.

La gestion des occlusions permet à notre système de détecter les événements suivants :

- Occlusion entre mains gauche et droite.
- Occlusion entre main gauche et visage.
- Occlusion entre main droite et visage.
- Occlusion entre les deux mains et le visage.

2.5 Détection et suivi du visage

Une fois que la segmentation en régions de l'image a été faite et que les occlusions entre les blobs d'intérêt sont détectées, on peut isoler les blobs qui potentiellement représentent une région de peau qui correspond à une partie du corps (Mains ou visage). Le but de cette partie est de pouvoir localiser et suivre le visage en le distinguant des mains. Pour se faire, on suppose comme dans [3] qu'une région R dans notre trame initiale représente le visage si :

- Le rapport entre la largeur et la hauteur d'un visage humain est inférieur à 2.25. Ceci revient à vérifier que $\frac{R_{\text{MajorAxisLength}}}{R_{\text{MinorAxisLength}}} < 2.25.$ (2.1)
- Le rapport entre l'aire et le carré du périmètre d'un visage humain est supérieur à 0.02. Ceci revient à vérifier que $\frac{R_{Area}}{(R_{Perimeter})^2} > 0.02.$ (2.2)

Le premier test permet de rejeter les régions qui sont trop longues et étroites pour être un visage, tel qu'un avant-bras. Le second est une mesure de circularité qui permet de cerner les régions de forme elliptique telle que la tête.

Dans [3], les auteurs ont aussi supposé que l'orientation du visage doit varier entre 45° et 135°. Cette hypothèse est plausible dans le cas où la tête de la personne reste droite durant une activité telle que la prise de médicaments. Dans notre cas, cette hypothèse est omise puisque nos algorithmes de suivi peuvent être utilisés pour la détection de n'importe quelle activité humaine. Quand la personne change l'orientation de son visage comme dans la figure 2.9 par exemple, la forme de ce dernier n'est plus elliptique. Aussi les mains dans certains cas peuvent avoir une forme elliptique et non étroite. Par conséquent on ne pouvait se fier à ces deux tests pour localiser et suivre le visage pour les autres trames des séquences présentées à notre système. Ces deux tests sont donc utilisés seulement pour la trame initiale.

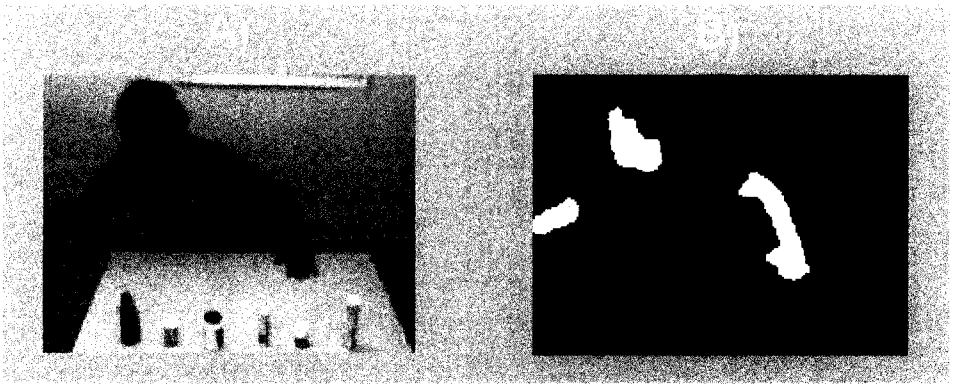

Figure 2.9 Exemple du visage ayant une forme non elliptique. A) Trames 8 de la séquence vidéo test, B) Détection des régions de la peau contenues dans cette trame.

Dans [43], les auteurs présentent les moments de Hu comme étant un descripteur de forme invariant. Ce dernier est obtenu à partir de quotients ou de puissances de moments. Un moment quant à lui représente une somme sur tous les pixels du modèle d'image pondéré par des polynômes liés aux positions des pixels. Considérons notre image *I(i,j)* après extraction des régions de peau. Le moment d'ordre $(p+q)$ pour chacun de nos blobs est défini comme

$$m_{pq} = \sum_i \sum_j i^p j^q I(i, j) \tag{2.3}$$

et le moment central d'ordre $(p+q)$ comme

$$\mu_{pq} = \sum_i \sum_j (i-\bar{i})^p (j-\bar{j})^q I(i,j) \tag{2.4}$$

Avec $\bar{i} = \frac{m_{10}}{m_{00}}$ (2.5) et $\bar{j} = \frac{m_{01}}{m_{00}}$ (2.6).

Le moment central normalisé d'ordre (p+q) est défini comme

$$\eta_{pq} = \frac{\mu_{pq}}{(\mu_{00})^{\lambda}} \tag{2.7}$$

avec

$$\lambda = \frac{p+q}{2} + 1. \tag{2.8}$$

Ces moments sont invariants par translation et changement d'échelle. Les moments de Hu sont déduits à partir des moments normalisés et sont invariants vis-à-vis les translations, rotation et changement d'échelle. L'équation 2.9 présente ces moments.

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{30})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(3\eta_{21} + \eta_{03})^2] \\
&+ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(3\eta_{30} - \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{2.9}$$

Dans notre méthode, une fois que la localisation du visage dans la première trame de notre séquence est faite, on calcule le moment de Hu d'ordre deux du blob correspondant et le suivi du visage est réalisé en comparant ce descripteur de forme. En effet, pour les trames qui suivent, on calcule les moments de Hu des régions de peau après segmentation puis on les compare au moment de la région correspondante au visage de la trame précédente. Pour justifier l'utilisation des moments de Hu, on a commencé par analyser

une séquence de trames dans laquelle le visage change d'orientations comme le montre la figure 2.10.

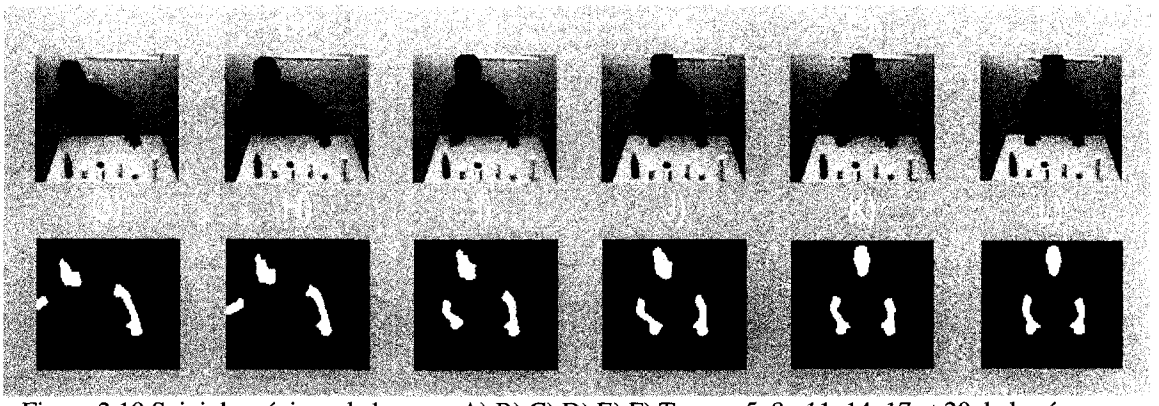

Figure 2.10 Suivi des régions de la peau. A) B) C) D) E) F) Trames 5, 8, 11, 14, 17 et 20 de la séquence vidéo test, G) H) I) J) K) L) Détection des régions de la peau contenues dans chacune des trames.

Par la suite, on a examiné dans un premier temps l'évolution des moments de Hu d'ordre 1(ϕ_1) et 2 (ϕ_2) des blobs correspondants au visage pour cette même séquence.

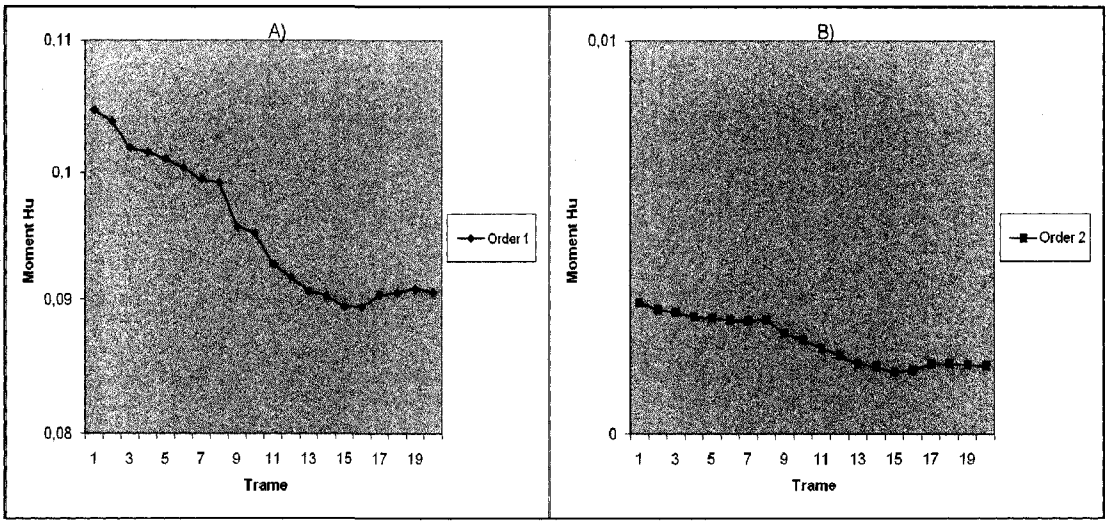

Figure 2.11 Comparaison de l'évolution des moments de Hu d'ordre1 et 2 du visage pour la séquence de la figure 2.10. A) Ordre1, B) Ordre2.

La figure 2.11 montre que les moments de Hu d'ordre 1 sont plus sensibles aux changements d'orientations du visage par rapport aux moments de Hu d'ordre 2. Ces derniers ont donné des résultats très satisfaisant et par conséquent c'est l'ordre 2 que nous avons utilisé pour le suivi du visage comme mentionné précédemment. On s'est arrêté à

l'ordre 2 puisque sa variation en valeur absolue pour le suivi du visage est trop petite comme le montre la figure précédente et aussi pour ne pas compliquer les calculs. En effet, comme le montrent les équations 2.9, à chaque fois qu'on augmente l'ordre de ce descripteur de forme, la complexité du programme augmente. L'idée de l'utilisation d'un descripteur de forme pour le suivi du visage est due au fait que la forme du visage est différente de celle des mains et varie moins lors d'une activité humaine. Cette considération est vérifiée par la figure 2.12. La section qui suit présente l'approche utilisée pour la localisation et le suivi des mains.

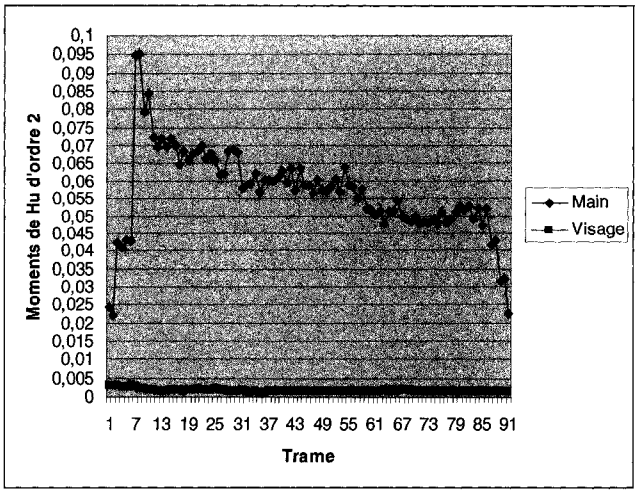

Figure 2.12 Comparaison de l'évolution des moments de Hu d'ordre 2 de la main et du visage pour une séquence de trames.

2.6 Détection et suivi des mains

Afin de pouvoir détecter l'activité humaine, la prise de médicaments dans notre cas, notre système doit pouvoir localiser et suivre les mains de l'utilisateur dans la séquence d'entrée. Pour ce faire, on suppose que les blobs restants après la localisation du blob représentant le visage correspondent aux mains. Pour distinguer les deux mains, on suppose que la main gauche se trouve toujours à gauche de la main droite et vice-versa, ce qui est vrai en général. Une analyse des blobs selon leur position horizontale nous permet de distinguer les deux mains. Dans notre application, la droite et la gauche sont définies selon le point de vue de la personne qui se trouve dans la séquence et non selon celui de la caméra. La complexité de la détection et du suivi de la main réside dans le cas

où la personne porte un chandail à manches courtes. Dans ce cas, il faut pouvoir localiser la main dans le bras. Pour notre application, on a conçu un algorithme qui se base sur les contours afin de permettre la détection de la main dans le bras. En effet, la main correspondra à la région englobée par le rectangle qui possède une densité des arêtes supérieure à cause des doigts.

On commence par transformer notre image binaire qui contient les régions de la peau en une autre image binaire qui détecte les contours de ces régions. Le but de la détection de contours est de repérer les points d'une image numérique qui correspondent à un changement brutal de l'intensité lumineuse. On a testé plusieurs méthodes permettant l'extraction des contours telles que les filtres de Prewitt, de Sobel et de Canny. La notion physique de filtre correspond à la notion mathématique de convolution. En effet, une convolution entre ces filtres et les trames d'une séquence vidéo suivie d'un seuillage de l'image résultante nous permet d'extraire les contours. Canny [44] a cherché à définir des critères afin d'obtenir un filtre optimal pour la détection de contour. Ces critères sont les suivants :

- Bonne détection : détecter un maximum de contours ;
- Bonne localisation : les points détectés doivent être les plus proches possibles du vrai contour ;
- Réponse unique : minimiser le nombre de contours détectés plusieurs fois.

Ces critères se traduisent par des conditions sur la réponse impulsionnelle du filtre et débouchent sur des détecteurs de contours très performants. L'avantage du filtre de Canny est qu'il considère non seulement l'intensité du gradient mais aussi sa direction. Par conséquent, il est possible d'éliminer un pixel qui pointe vers deux pixels de valeur supérieure car ce n'est pas un maximum local. Il faut ensuite effectuer un seuillage par hystérésis. Pour cela on fixe deux seuils, un seuil haut *sh* et un seuil bas *sb*. On commence par sélectionner les points qui dépassent le seuil haut et on applique ensuite le seuil bas en ne conservant que les composantes connexes qui contiennent un point au

dessus de sh. En d'autres termes, à partir de chaque point au dessus de *sh* on "suit" un chemin constitué de points au dessus de *sb*, ce chemin est le contour recherché.

La figure 2.13 montre le résultat obtenu pour ces trois méthodes pour une série de trames d'une séquence vidéo test. Les paramètres utilisés sont ceux choisis automatiquement par la fonction *edge* de Matlab.

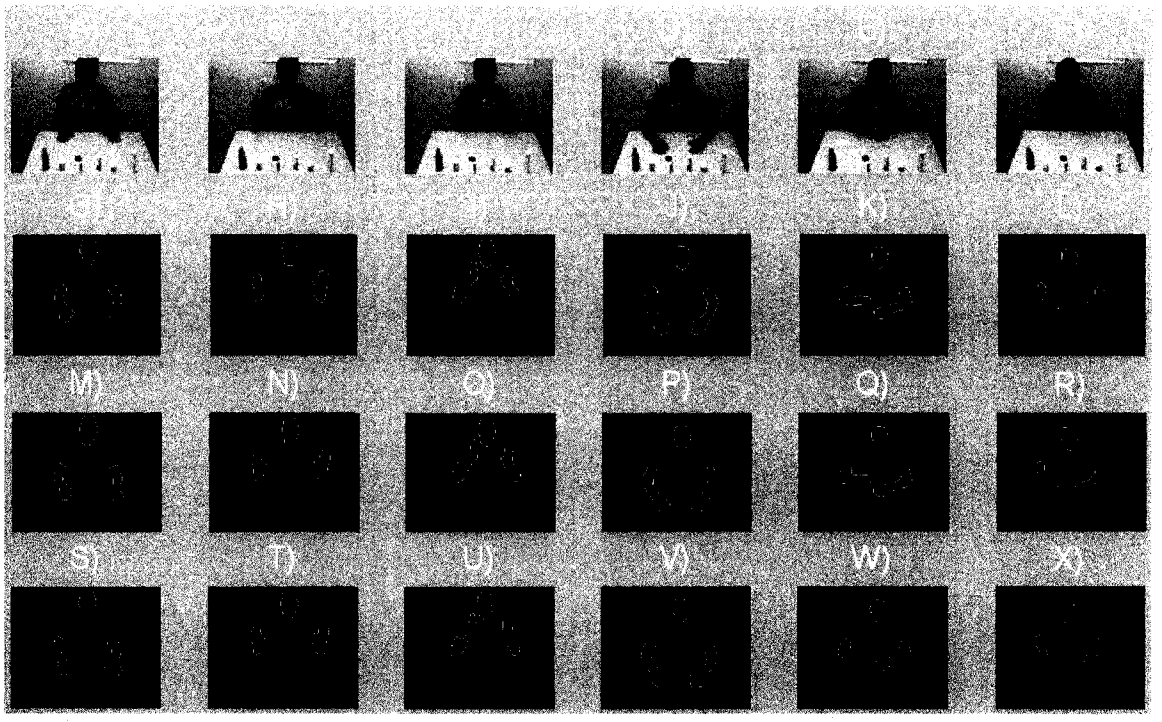

Figure 2.13 Détection des contours des régions de la peau. A) B) C) D) E) F) Trames 25, 125, 230, 325, 425 et 650 de la séquence vidéo test, G) H) I) J) K) L) Détection des contours avec la méthode Sobel, M) N) O) P) Q) R) Détection des contours avec la méthode Canny, S) T) U) V) W) X) Détection des contours avec la méthode Prewitt.

On peut constater qu'il n'y a pas une grande différence entre les trois méthodes d'extraction des arêtes. On remarque aussi que le contour n'est pas parfaitement défini ce qui pourrait induire le système en erreur. Pour pallier à ce problème on a utilisé des opérations de la morphologie mathématique. En effet, une fois que l'image du contour est extraite, on lui fait subir une opération de morphologie mathématique qui est dans notre cas une dilatation (augmentation de la taille des arêtes avec la fonction *imdilate* de MATLAB) afin de mieux définir les contours des régions contenants les mains de la

personne. Le résultat obtenu avec une détection de contour utilisant la méthode de Canny suivie d'une dilatation est présenté dans la figure 2.14.

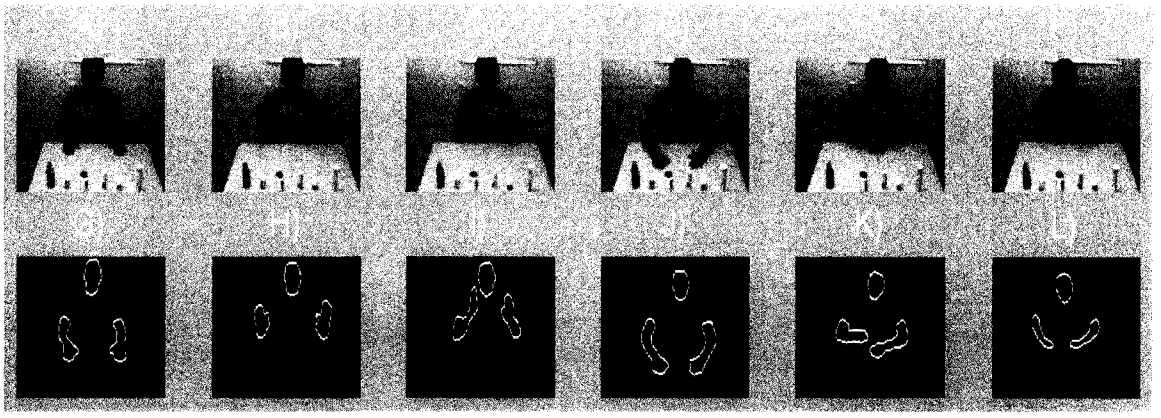

Figure 2.14 Détection des contours des régions de la peau. A) B) C) D) E) F) Trames 25, 125, 230, 325, 425 et 650 de la séquence vidéo test, G) H) I) J) K) L) Détection des contours avec la méthode Canny suivi d'une dilatation.

Une fois que les contours des régions de la peau sont bien tracés et afin de pouvoir suivre les mains, on définit notre modèle pour chacune des régions de peau qui ne représentent pas le visage. Notre modèle rectangulaire de la main servant à conserver la position de cette dernière pour chacune des trames de nos séquences vidéo est définie comme suit :

$$H = [B, L, C, F] \tag{2.10}$$

Avec B qui représente le plus petit rectangle qui englobe la région du bras (*Bounding Box*), C qui est le centre du rectangle représentant la main, L qui représente la largeur de la main et finalement F qui représente la densité des arêtes de la main. Pour chacune des trames, on suppose que la largeur de la main représente 4/5 de celle du visage. Cette proportion est déterminée expérimentalement et s'ajustera automatiquement si l'utilisateur change sa position par rapport à la caméra puisque la région du visage changera aussi de grandeur.

Ainsi, pour chacune des régions de la peau qui ne représente pas le visage, on commence par localiser les deux extrémités de ces régions qui peuvent englober la main. Cette localisation est effectuée en utilisant l'orientation du bras (verticale ou horizontale), les tailles du rectangle englobant et la largeur de la main calculée selon l'hypothèse précédente. La figure 2.15 montre deux exemples de détection des deux régions de peau

qui représentent les parties les plus probable de contenir la main avec deux orientations différentes du bras.

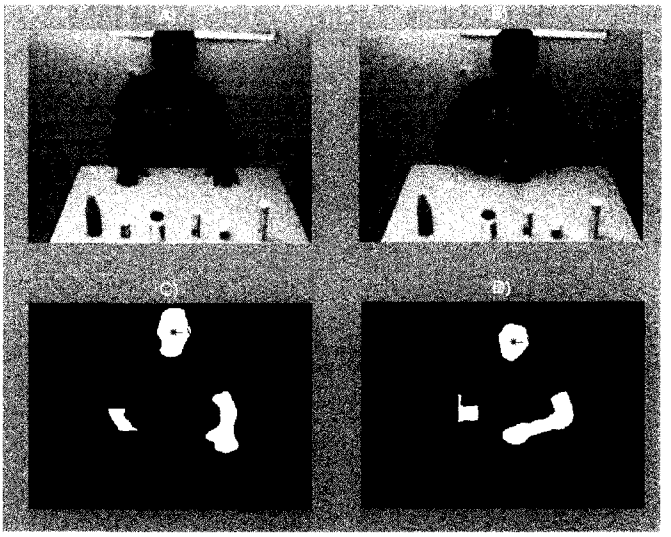

Figure 2.15 Détection des région qui peuvent contenir la main dans le bras droit A) B) Trames 25 et 425 de la séquence vidéo test, C) D) Détection des régions de la peau pouvant englobées la main.

Une fois les deux rectangles les plus probables de contenir la main localisés, on suppose pour la trame courante que la main contient plus d'arêtes à cause des doigts. On calcule donc le nombre de pixels des arêtes dans l'image contenant les contours. Alors, pour chacun des deux rectangles, on calcule la densité des arêtes comme définie en [26]. Le nombre de pixels arêtes dans chacun des rectangles probables de contenir la main est calculé selon l'équation 2.11.

$$F = \frac{\sum I_c(i,j) \mid I_c(i,j) == 1}{N} \qquad (2.11)$$

Avec N le nombre de pixels dans chacun des rectangles et I_c l'image contenant les arêtes des régions de la peau. Cette caractéristique de texture nous permettra de localiser la main. Finalement, on complète notre modèle de la main en calculant le centre du rectangle. Le suivi pour les cinq trames suivantes se fait par comparaison de la position des deux rectangles extraits pouvant englober la main de la trame courante avec la position du modèle précédent. Nous avons supposé que pour cinq trames la main ne

change pas assez de position, ce qui minimise les erreurs du suivi par la comparaison du modèle de centres. En cas d'erreurs, la trame suivante qui va calculer la densité d'arêtes va permettre au système de se rattraper et de réinitialiser la bonne position de la main. Comme mentionné précédemment, pour valider nos algorithme de détection des régions de peau et du suivi du visage et de la main, on a choisi la détection l'activité de prise de médicaments. Pour ce faire, notre système doit détecter et identifier les bouteilles de médicaments présentes dans la séquence vidéo d'entrée. La section qui suit présente l'approche utilisée pour la localisation et le suivi de ces objets d'intérêts.

2.7 Détection et suivi des bouteilles de médicaments

Pour compléter le suivi des objets intervenants dans l'activité de prise de médicaments, on a utilisé des bouteilles de médicaments de formes et couleurs différentes. Pour créer notre base de données contenant les informations concernant la couleur et la forme des bouteilles de médicaments, un ensemble d'images comportant des régions complètes de chacune de ces bouteilles est présenté au système. Les histogrammes de couleurs ainsi que les moments de Hu d'ordre deux de ces régions sont calculés. La couleur de la table où seront posées les bouteilles est également présentée au système. Ceci évitera les recherches multi-échelles dans toute l'image et diminuera ainsi les fausses identifications et détections des bouteilles. En effet, la caméra permettant la prise de nos séquences vidéo est placée en face de la personne et légèrement au-dessus de sa tête comme montré à la section 2.2. Cet emplacement de la caméra nous permet de diminuer les possibilités de détection de faux contacts entre les mains et les bouteilles de médicaments et de pouvoir utiliser la table comme arrière-plan des objets qui s'y trouvent. Les seuils utilisés lors de notre travail pour localiser la table sur laquelle les bouteilles sont posées ont été déterminés expérimentalement et sont illustrés au tableau 2.2.

Tableau 2.2 Seuils utilisés avec l'espace de couleur *HSV* afin de détecter les pixels de la table.

Canal	Seuil inférieur	Seuil supérieur
Hue ou Couleur pure (H)	0.1	0.3
Saturation (S)	0.2	0.6
Valeur (V)	0.85	1

La figure 2.16 illustre le résultat du seuillage et de la segmentation de la région contenant la table pour une trame d'une séquence vidéo test avec les mêmes techniques utilisées dans la section 2.3 pour la détection de la peau mais cette fois-ci avec des seuils différents.

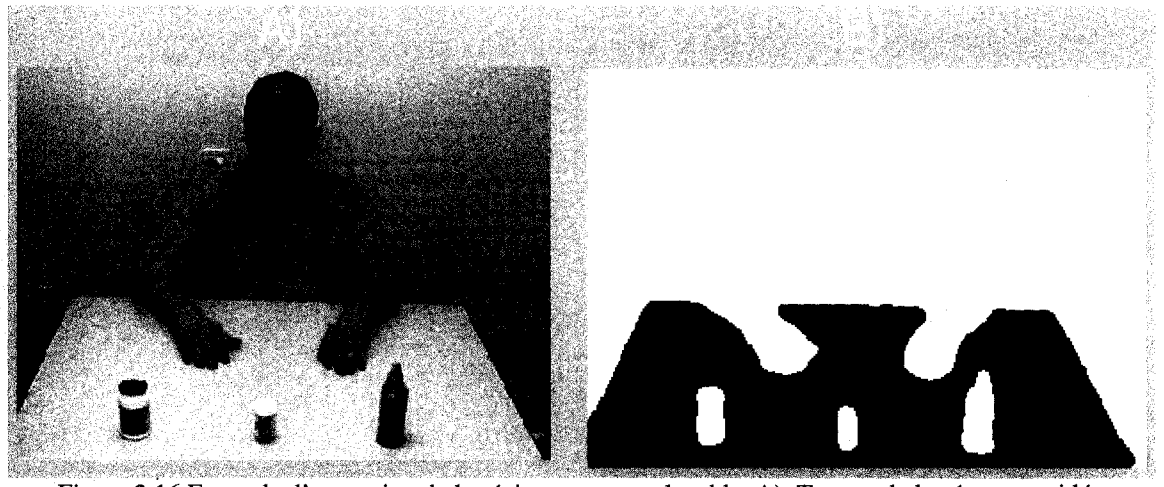

Figure 2.16 Exemple d'extraction de la région contenant la table. A), Trames de la séquence vidéo originale, B), Détection de la table et des objets qui s'y trouvent.

Pour tous les objets se trouvant sur la table, on calcule leurs histogrammes de couleurs ainsi que leurs moments de Hu d'ordre 2 présentés précédemment. Dans la première trame, on détecte les objets qui ont les mêmes caractéristiques d'apparence (couleur + forme) que celles des bouteilles de médicaments présentées au système. Comme défini dans [24], les histogrammes de couleurs sont considérés comme des vecteurs et la

distance entre eux est calculée en utilisant le prix minimum pour passer d'une distribution à l'autre (MDPA) selon l'équation 2.12.

$$D(h(I), h(M)) = \sum_{j=0}^{K-1} \left| \sum_{k=0}^{j} (h(I)[k] - h(M)[k]) \right|. \qquad (2.12)$$

La figure 2.17 présente un exemple de détection et d'identification des bouteilles de médicaments utilisées dans des séquences de prises de médicaments testées.

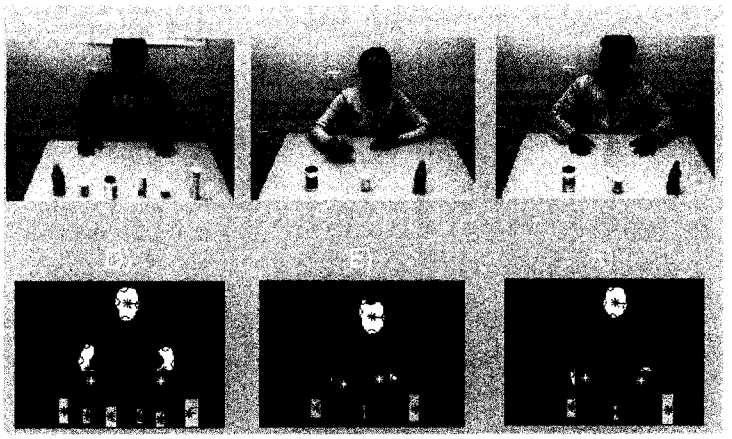

Figure 2.17 Exemple de détection des régions de la peau et des bouteilles de médicaments. A) B) C) Trames des séquences vidéos originales, D) E) F) Détection et suivi des objets concernés.

La figure montre que les bouteilles de médicaments sont bien identifiées et que notre système ne les confond pas avec les mains. Étant donné que les bouteilles de médicaments sont des objets rigides, on compare les centroïdes des objets détectés sur la table de la trame courante avec les centroïdes des bouteilles de médicaments détectées dans la trame précédente pour pouvoir les identifier dans toutes les trames de notre séquence source.

On suppose qu'au début de la séquence les bouteilles ne sont pas en occlusion entre elles et avec les mains. Ceci nous permet de connaitre le nombre de bouteilles initiales présentes dans la séquence vidéo. Notre système détecte l'événement « Prise d'une bouteille de médicament » lorsque celle-ci n'est plus déposée sur la table. Cette condition se vérifie quand le nombre de bouteilles détectées dans la trame courante est inférieur à

celui de la trame précédente. On calcule donc les distances entre les centroïdes des deux mains et le centroïde de la bouteille prise pour déterminer la main qui manipule la bouteille de médicaments. Les événements « Main gauche manipule la bouteille » ou « Main droite manipule la bouteille » sont alors détectés.

Afin de limiter les fausses détections et erreurs de suivi de la bouteille lorsqu'elle est en occlusion ou en contact avec les mains, on a utilisé l'approche MS présenté à la section 2.4. En effet, le suivi de la bouteille de médicament dans notre cas est suspendu pendant la durée de l'occultation ou le contact. Lorsque l'objet est redéposé sur la table, l'événement « Dépôt de la bouteille » est détecté et l'identification des objets présents sur la table se poursuit. L'algorithme utilisé pour la reconnaissance de l'activité humaine est présenté dans la section suivante.

2.8 La reconnaissance de l'activité humaine

Pour la détection de l'activité humaine, qui est dans le cadre de notre étude la prise de médicament, on a utilisé un réseau de Petri. Les réseaux de Petri furent inventés en 1964 par Carl Adam Petri. Ces derniers se représentent par un graphe composé de deux types de nœuds (places et transitions) reliés par des arcs. Un réseau de Petri évolue lorsqu'on exécute une transition : des jetons sont pris dans les places en entrée de cette transition et envoyés dans les places de sortie. On reconnait donc un scénario lorsqu'un jeton est placé à la suite du dernier événement de la séquence. Dans [45], les auteurs ont évoqué les avantages de l'utilisation des réseaux de Petri pour la représentation et la reconnaissance des événements. Parmi ces avantages, on trouve la représentation des événements de façon séquentielle, simultanée et synchronisée. Notre réseau de Petri a été conçu pour détecter tous les scénarios possibles de la prise de médicaments selon les contraintes imposés dans ce projet. Il possède sept places et dix transitions comme le montre la figure 2.18.

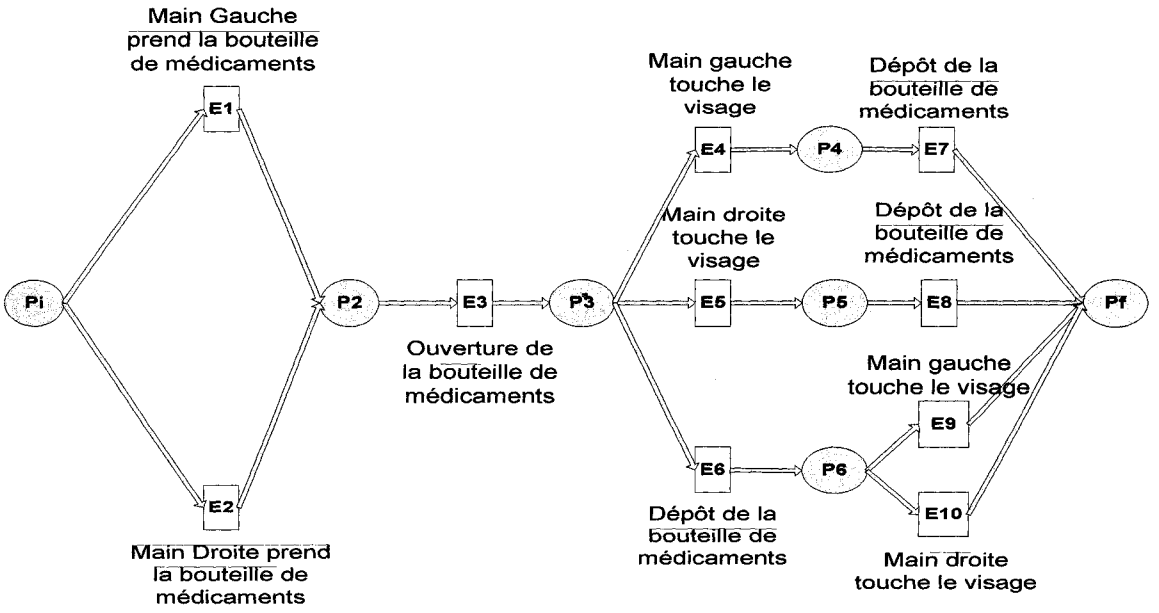

Figure 2.18 Réseau de Petri utilisé pour la reconnaissance de la prise de médicaments.

Au départ le jeton est déposé dans la place initiale Pi. Si un des événements E1 ou E2 se produit on déplace le jeton à la place P2. On utilise les relations logiques définies en [45] pour la construction de notre réseau de Petri. Un exemple de relations logiques est illustré à la figure 2.19. On peut voir sur notre réseau que le jeton sera passé de la place Pi à la place P2 si un des événements E1 ou E2 est détecté.

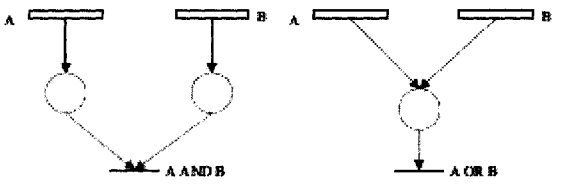

Figure 2.19 Exemple de relations logiques. Figure extraite de [45].

L'activité de prise de médicaments est reconnue lorsque le jeton atteint la place Pf. On note que pour l'événement E3, on suppose que l'ouverture de la bouteille se produit lorsque la main de l'usager manipule celle-ci et qu'il y a occlusion ou contact entre les deux mains. Les relations entre les transitions du réseau de Petri et les autres événements détectés par notre système sont comme suit :

- E1 : Main gauche manipule la bouteille.
- E2 : Main droite manipule la bouteille.
- E3 : Occlusion entre main gauche et droite + Prise d'une bouteille de médicament.
- E4 et E9 : Occlusion entre main gauche et visage.
- E5 et E10 : Occlusion entre main droite et visage.
- E6, E7 et E8 : Dépôt de la bouteille.

Pour valider les transitions et éviter les fausses détections des événements, ces derniers ne sont validés que si leurs durées dépassent un certain nombre de trames. Le nombre de trames minimal (NTM) que doit durer chacun des événements en utilisant une caméra avec un taux de 7.5 frames par seconde est déterminé expérimentalement et il est présenté dans le tableau 2.3.

Tableau 2.3 Le nombre minimum de trames que doit durer les événements utilisés dans le réseau de Petri avec caméra à 7.5 trames/seconde.

Événements	E1 E2	E3	E4 E5 E9 E10	E6 E7 E8
NTM	5	10	10	1

CHAPITRE 3 RÉSULTATS ET DISCUSSION

Dans cette section on présentera les résultats obtenus sur des séquences vidéo avec une fréquence d'acquisition de 7.5 images par seconde et une résolution de 320x240 prises par une caméra Sony DFW-SX910 (voir figure 3.1). L'application a été programmée avec Matlab. Ce dernier est un langage de programmation de haut niveau pour le calcul numérique. Il est particulièrement performant pour le calcul matriciel, car sa structure de données est basée sur les matrices et il dispose de possibilités d'affichage très riches. Il s'agit d'un langage qui permet un développement très rapide mais qui a l'inconvénient de ne pas avoir un temps d'exécution aussi rapide qu'un langage comme C. En plus de fonctions de bases pour le calcul matriciel, Matlab dispose de nombreuses librairies de fonctions spécialisées appelées « toolbox » dans différents domaines. Celle qui nous intéresse le plus particulièrement est la librairie « traitement d'image » qui possède des fonctions qui nous ont été utiles pour la réalisation du projet. Parmi ces fonctions, on trouve les fonctions de :

- Calcul des histogrammes de couleurs,
- Détection de contours et
- Binarisation et morphologie mathématique.

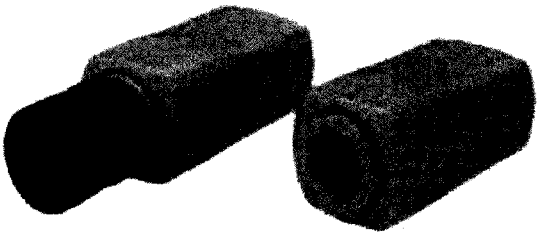

Figure 3.1 Caméra Sony DFW-SX910.

Ce chapitre est divisé en quatre grandes sections : les méthodologies expérimentales ainsi que les résultats comprenant une discussion de la partie détection des régions de la peau (section 3.1), les méthodologies expérimentales ainsi que les résultats comprenant une discussion de la partie détection et suivi du visage et des mains (section 3.2), les méthodologies expérimentales ainsi que les résultats comprenant une discussion de la partie reconnaissance de la prise de médicaments (section 3.3) et finalement une section 3.4 portant sur le temps d'exécution de notre système.

3.1 Détection des régions de la peau

3.1.1 Méthodologie expérimentale

Comme mentionné précédemment, pour pouvoir localiser et suivre le visage et les mains, on s'est basé sur la couleur de la peau. Notre algorithme de seuillage et de segmentation présenté dans la section 2.3 a été testé sur différentes personnes avec des images de différentes résolutions. En effet, pour évaluer la performance de la détection des régions de peau, nous avons utilisé 15 séquences vidéo prises au LITIV dont 12 représentent une activité de prise de médicaments et des images prises au LITIV et d'autres provenant du web. Nous avons ensuite appliqué nos algorithmes sur ces images et sur chaque trame des séquences vidéo et avons vérifié si la détection était correcte. Pour la peau, une détection est correcte si les régions de peau détectées avec notre algorithme correspondent aux régions de peau contenues dans l'image testées. L'efficacité de notre algorithme pour la détection des régions de la peau a été mesurée selon la métrique suivante : *Efficacité=(Total des images avec détection correcte de la peau / Total des images testées).*

3.1.2 Résultats

La présente section montre les résultats des expérimentations et des tests effectués afin de valider nos algorithmes de détection de la peau. Des exemples d'extraction des régions de peau contenues dans quelques-unes des images prises au LITIV et du web sont présentés à la figure 3.2. Cette figure montre les résultats obtenus à partir d'images sources représentant des cas particuliers qui peuvent engendrer des erreurs de détection des

régions de peau. Les images de la figure 3.2 confirment que les seuils qu'on a choisi provoquent peu de conflits avec des couleurs différentes que celles de la peau, peuvent détecter différents types de peau dans différentes conditions d'illumination et permettent de localiser le visage malgré le port de lunette ou casquette. Les images présentées dans cette figure appuient les recherches antérieures dans le fait que l'espace de couleur HSV avec les seuils présentés permettent de détecter différents types de peau.

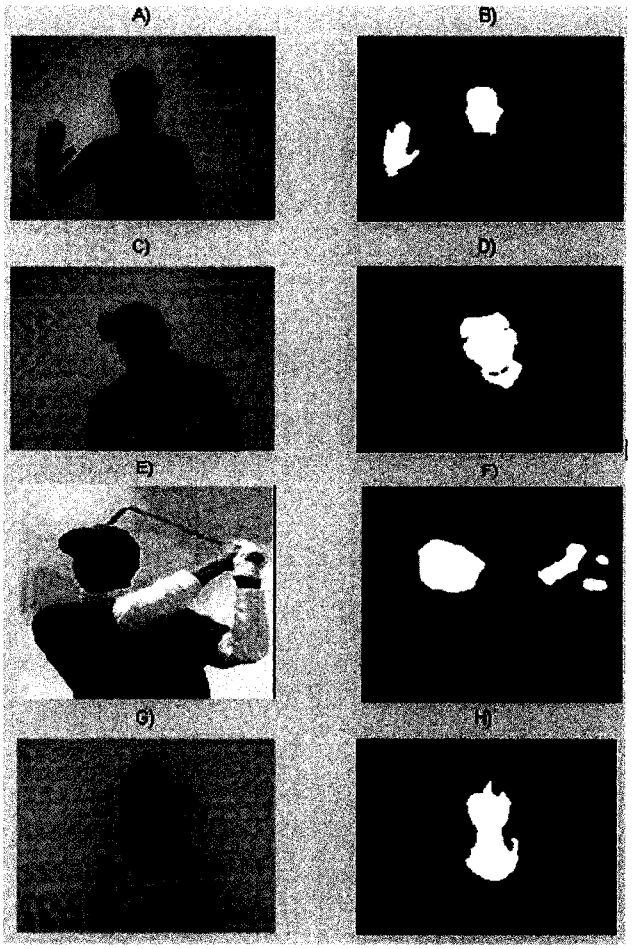

Figure 3.2 Exemple d'extraction des régions de la peau. A), C), E), G), Images sources, B), D), F), H), Détection des régions de la peau contenues dans chacune des images selon l'algorithme de seuillage et de segmentation présenté à la section 2.3.

En effet, on constate que les régions de peau noire de l'image E ont était bien trouvées. Cependant, dans cette image, on voit qu'un petit bout du bâton de golf a été classifié comme étant une région de peau. C'est pour cela qu'on a enlevé tous les objets de

couleurs semblables à celles de la peau de l'environnement où les médicaments sont pris. L'image C montre que le visage est détecté malgré le port de casquette et de lunette. Pour les images prises au LITIV et du web, notre algorithme a permis la détection des régions de peau sur différentes personnes avec une efficacité de 68% (voir tableau 3.1). Aussi, notre algorithme a permis la détection des régions de peau sur différentes séquences vidéo prises au LITIV avec une efficacité de 100%. Par conséquent l'algorithme de seuillage et de segmentation utilisé afin d'extraire les pixels de la peau a démontré son efficacité pour permettre la détection et le suivi du visage et des mains et la reconnaissance de l'activité humaine qui est dans notre cas la prise de médicaments.

Tableau 3.1 Résultats de la détection des régions de la peau.

	Nombre d'images dans la séquence	Images avec détection correcte des régions de la peau
Séquences prises au LITIV avec prise de médicaments	2937	2937
Séquences prises au LITIV	405	405
Total des trames des séquences testées	3342	3342
Efficacité		**100%**
Images prises au LITIV	15	12
Images prises du web	10	5
Total des images testées	25	17
Efficacité		**68%**
Total des images et trames testées	3367	3359
Efficacité		**99%**

3.2 Détection et suivi du visage et des mains

3.2.1 Méthodologie expérimentale

Une fois les parties du corps détectées, notre système procède à la détection et le suivi de ces parties ainsi qu'à l'identification et la localisation des bouteilles de médicaments. Pour comparer les performances de nos algorithmes de détection et de suivi des mains et du visage par rapport à des méthodes existantes, le système a été initialement examiné sur quatre séquences avec prise de médicaments dans des conditions différentes sur un ensemble de 3 utilisateurs comme illustré à la figure 3.3.

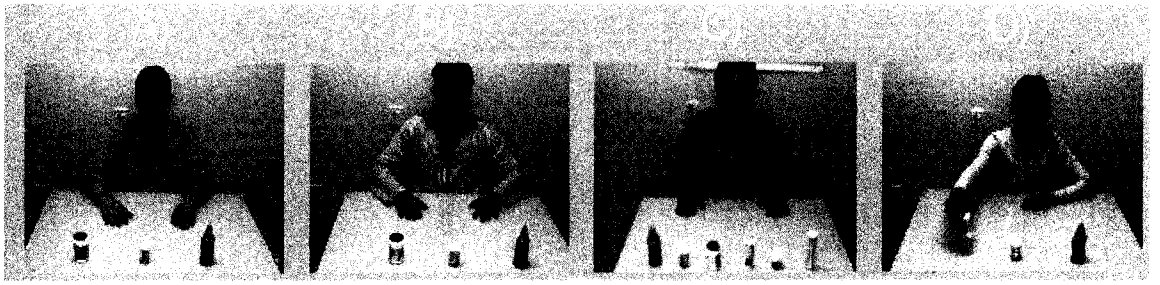

Figure 3.3 Séquences vidéo utilisée pour évaluer la performance de nos algorithmes. A), Séquence François B), Séquence Soufiane1 C), Séquence Soufiane2 D), Séquence Atousa.

Pour trouver le nombre de trames avec suivi correct du visage et des mains avec les méthodes testées, on a étiqueté pour chacune des trames les parties du corps localisées et suivies par notre système et on a comparé ces étiquettes avec la position réelle «ground-truth» du visage et des mains. L'efficacité de notre algorithme pour la détection des régions de la peau a été mesurée selon la métrique suivante : *Efficacité=(Total des trames avec suivi correct / Total des images testées).*

Dans un premier temps, on a commencé par comparer l'efficacité de notre algorithme avec celle du Mean-shift. Pour ce faire, il fallait sélectionner les régions qu'on veut suivre, dans notre cas le visage et les deux mains, tout en entrant la taille, en pixels, qu'aurait le carré qui englobera ces régions durant le suivi. Pour une main, la sélection appropriée est entre 25 et 35 pixels et pour le visage, elle devra être entre 45 et 55. Par la suite on appelle la fonction qui fait le suivi avec la méthode Mean-shift telle que décrite dans la section 1.2.3.2.1. Dans un second temps, on a évalué l'efficacité du filtrage

particulaire pour le suivi des parties du corps d'intérêt. Pour ce faire, il fallait sélectionner les régions qu'on voulait suivre, dans notre cas le visage et les deux mains, tout en entrant la taille, en pixels, qu'aurait le carré qui englobera ces régions durant le suivi. Pour une main, la sélection appropriée est entre 25 et 35 pixels et pour le visage, elle devra être entre 45 et 55 pixels. Par la suite, on entre le nombre de particules que l'on souhaite utiliser et on appelle la fonction qui fait le suivi avec la méthode du filtrage particulaire telle que décrite dans la section 1.2.3.2.2. Le nombre de particules utilisées dans le cadre de nos tests est 50. Plus ce nombre est grand, plus le suivi des régions sera précis et par conséquent plus le script sera lent.

D'autres séquences de prise de médicaments ont été utilisées pour évaluer la performance de nos algorithmes de localisation et de suivi des parties du corps sur différentes personnes. La figure 3.4 illustre quelques-unes de ces séquences.

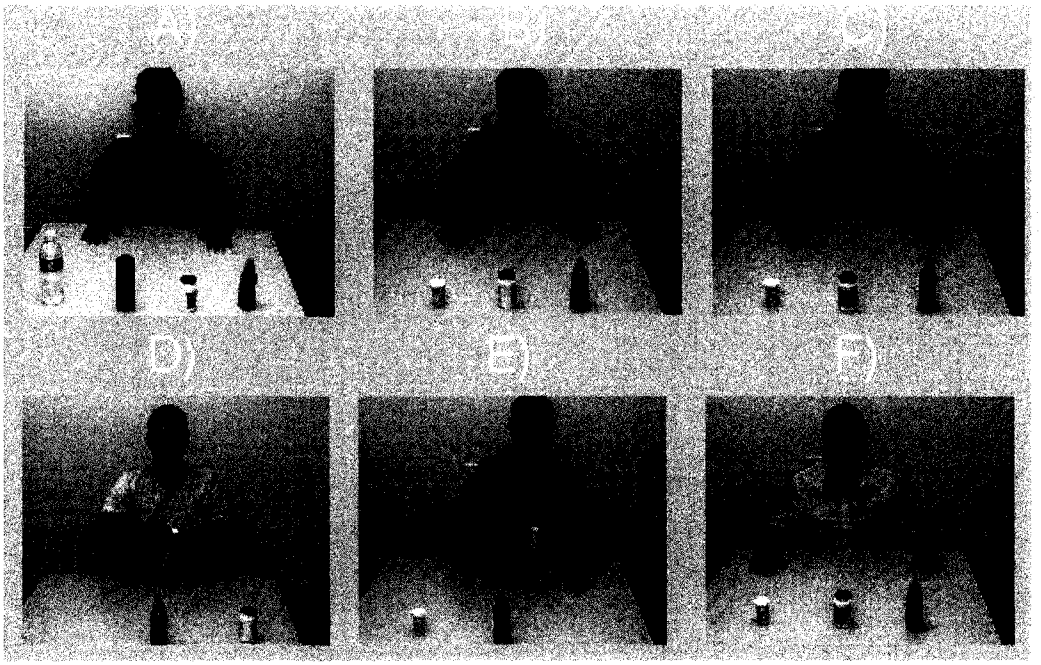

Figure 3.4 Séquences vidéo utilisée pour évaluer la performance de nos algorithmes. A), Séquence Soufiane3 B), Séquence Karim1 C), Séquence Karim2 D), Séquence Ali1 E), Séquence Ali2 F), Séquence Younes1.

3.2.2 Résultats

La présente section montre les résultats des expérimentations et des tests effectués afin de valider nos algorithmes de détection et de suivi des objets d'intérêt (visage, mains et bouteilles de médicaments). Un exemple de suivi dans une séquence vidéo présentant une prise de médicaments est présenté à la figure 3.5.

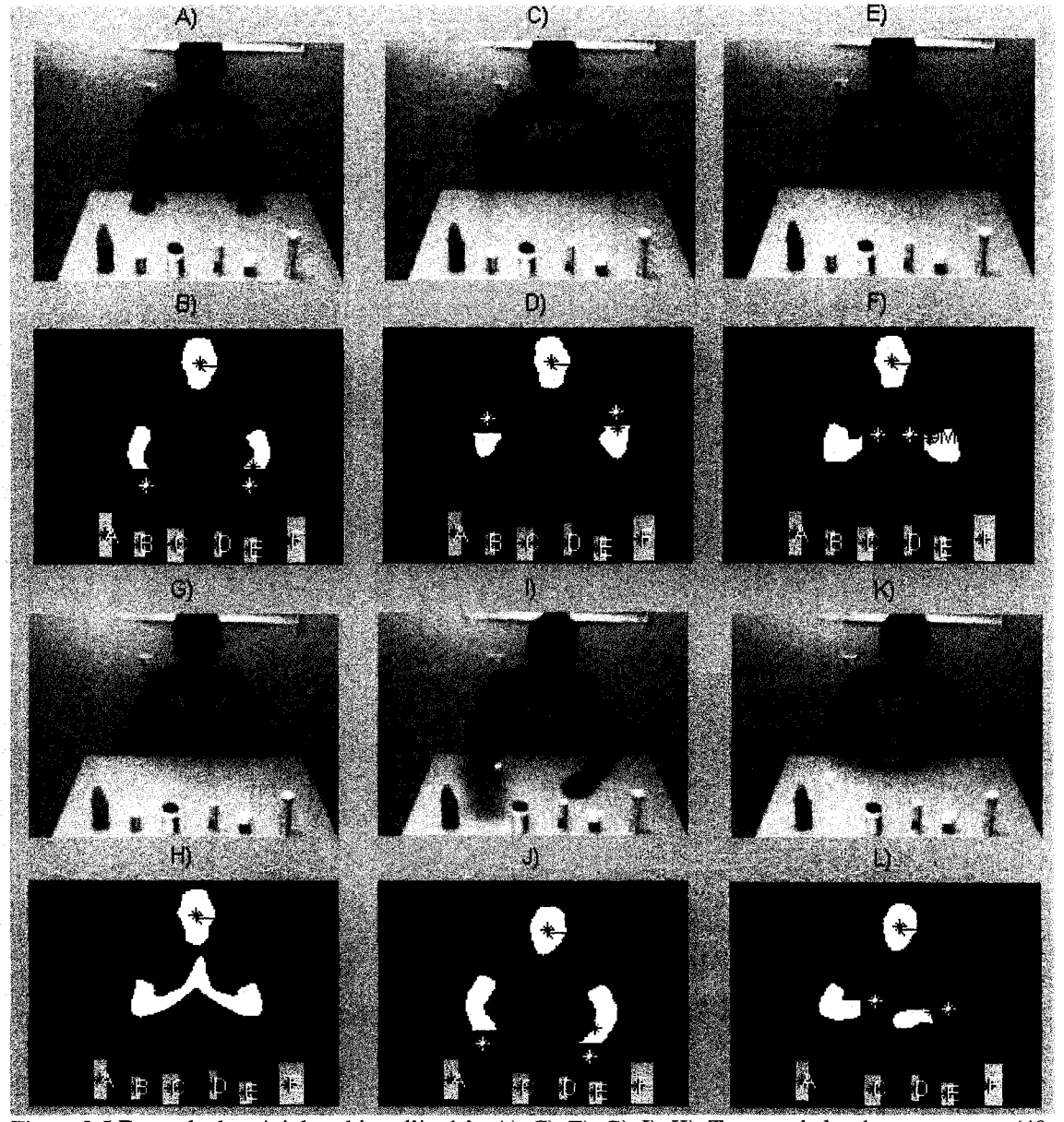

Figure 3.5 Exemple de suivi des objets d'intérêt. A), C), E), G), I), K), Trames de la séquence source (40, 125,170,190, 350 et 480), B), D), F), H), J), L), Localisation et suivi du visage, des mains et des bouteilles de médicaments.

On a commencé par comparer les performances de nos algorithmes de suivi du visage et des mains avec des méthodes existantes. Les résultats suite à l'application de nos algorithmes sur l'ensemble des séquences vidéo de la figure 3.3 sont présentés au tableau 3.2.

Tableau 3.2 Résultats de la détection et du suivi des parties du corps pour 4 séquences vidéo.

	Nombre de trames dans la séquence	Trames avec suivi correct du visage	Trames avec suivi correct des mains
Séquence François	110	107	100
Séquence Soufiane1	154	150	152
Séquence Soufiane2	694	691	674
Séquence Atousa	145	143	144
Total	1103	1091	1070
Efficacité (%)		**99%**	**97%**

Le tableau 3.2 montre que les méthodes élaborées permettent la localisation et le suivi du visage avec une efficacité de 99% pour les séquences testées. Celles utilisées pour la détection et le suivi des mains présentent une efficacité de 97% sur l'ensemble des séquences testées. Cependant, dans des cas particuliers dans des séquences vidéo, la main possède une forme qui s'approche à celle du visage. Dans ce cas, les moments de Hu d'ordre 2 sont presque égaux et le système ne peut pas classer la bonne région du visage et, par conséquent, il commet une erreur dans la localisation des mains. Ceci explique la plus faible précision du système pour le suivi des mains par rapport au visage. Lorsque la personne porte un chandail à manche courte (figure 3.5K), parfois la main possède une densité d'arrête inférieure à celle de l'autre extrémité du bras. Cela représente une autre source d'erreurs pour notre système dans le suivi des mains. On note que dans le cas où il y a occultation ou contact entre les parties du corps, le suivi des mains et du visage s'arrête comme expliqué dans la section 2.4 et comme illustré dans la figure 3.5H. Si

l'événement de fusion des blobs a été correctement détecté, on compte les trames pour la durée de l'occlusion comme étant des trames avec un suivi correct des parties du corps qui se sont fusionnées. Autre limite du système, c'est quand la personne porte une montre comme illustré à la figure 3.6. Le système ne segmente pas correctement les blobs des deux bras à cause de la couleur de la montre et par conséquent la localisation et le suivi des mains ne peuvent se faire correctement. En effet, dans ce cas l'événement d'occlusion entre main gauche et droite (figure 3.6B) ne sera pas détecté à cause de la présence de la montre.

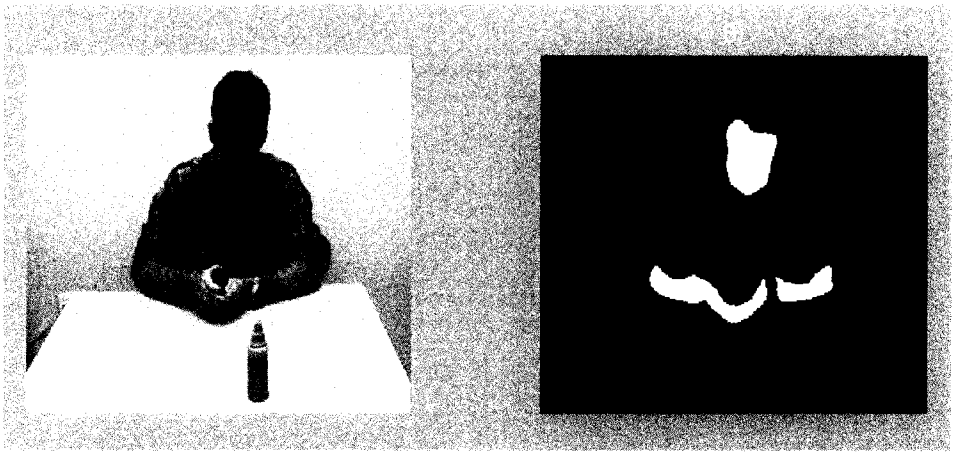

Figure 3.6 Exemple d'extraction des régions de la peau. A), Image source, B), Détection des régions de la peau contenues dans chacune des images

Pour mieux évaluer la performance de nos méthodes du suivi du visage et des mains, on les a comparées avec les méthodes du décalage moyen (Mean-shift) et du filtrage particulaire présentées dans la section 1.2.3.2 en les appliquant sur les mêmes séquences vidéo présentées à la figure 3.3. Les résultats de cet algorithme pour l'ensemble des séquences vidéo sont présentés au tableau 3.3. Ce dernier montre que la méthode Mean-shift permet le suivi du visage avec une efficacité de 96% et que le suivi des mains avec cette méthode présente une efficacité de 80% sur l'ensemble des séquences testées.

Tableau 3.3 Résultats de la détection et du suivi des parties du corps pour 4 séquences vidéo avec la méthode du décalage moyen (Mean-shift).

	Nombre de trames dans la séquence	Trames avec suivi correct du visage	Trames avec suivi correct des mains
Séquence François	110	102	79
Séquence Soufiane1	154	153	119
Séquence Soufiane2	694	658	602
Séquence Atousa	145	141	77
Total	1103	1054	862
Efficacité (%)		**96%**	**80%**

Par la suite, on a évalué l'efficacité du filtrage particulaire pour le suivi des parties du corps d'intérêt. Le résultat de cet algorithme pour l'ensemble des séquences vidéo est présenté au tableau 3.4.

Tableau 3.4 Résultats de la détection et du suivi des parties du corps pour 4 séquences vidéo avec la méthode du filtrage particulaire.

	Nombre de trames dans la séquence	Trames avec suivi correct du visage	Trames avec suivi correct des mains
Séquence François	110	92	55
Séquence Soufiane1	154	136	93
Séquence Soufiane2	694	561	424
Séquence Atousa	145	123	49
Total	1103	912	621
Efficacité (%)		**83%**	**56%**

Le tableau 3.4 montre que la méthode du filtrage particulaire permet le suivi du visage avec une efficacité de 83% et que le suivi des mains avec cette méthode présente une efficacité de 56% sur l'ensemble des séquences testées. On peut clairement remarquer que nos méthodes basées sur les approches par apparence sont plus efficace que les autres approches testées.

La limitation principale des approches prédictives est qu'elles sont moins fiables lorsque des changements brusques de direction des objets suivis se produisent. Dans ces cas, les prédictions sont souvent fausses comme dans le cas du filtrage particulaire qui a eu seulement une efficacité 56% pour le suivi des mains qui changent beaucoup de position et de direction dans les séquences testées. Le problème avec cette méthode c'est qu'une fois les parties du corps suivies s'occultent ou se touchent, l'algorithme perd la position de l'objet suivi et ceci jusqu'à la fin de la séquence ce qui diminue grandement l'efficacité du suivi.

D'autres séquences de prise de médicaments (voir figure 3.4) ont été utilisées pour évaluer la performance de nos algorithmes de localisation et de suivi des parties du corps sur différentes personnes. Les résultats suite à l'application de nos algorithmes sur l'ensemble de ces séquences vidéo sont présentés au tableau 3.5. On note qu'on a utilisé la même technique pour l'évaluation de la performance et que les méthodes élaborées dans le cadre de notre recherche permettent la localisation et le suivi du visage avec une efficacité de 99% pour les séquences testées. Celles utilisées pour la détection et le suivi des mains présentent une efficacité de 96% sur l'ensemble de ces six séquences testées.

Tableau 3.5 Résultats de la détection et du suivi des parties du corps pour les 6 séquences vidéo illustrées à la figure 3.4.

	Nombre de trames dans la séquence	Trames avec suivi correct du visage	Trames avec suivi correct des mains
Séquence Soufiane3	321	311	300
Séquence Karim1	140	137	130
Séquence Karim2	229	229	229
Séquence Ali1	131	131	125
Séquence Ali2	376	376	376
Séquence Younes1	237	230	220
Total	1434	1414	1380
Efficacité (%)		**99%**	**96%**

Les tableaux 3.2 et 3.5 montrent que l'efficacité de notre système est plus élevée lorsque la personne porte un chandail à manches longues. En effet, le tableau 3.6 prouve clairement cette constatation et montre que l'efficacité de nos algorithmes de détection et de suivi des mains pour les séquences testées est de 99% dans le cas ou la personne porte un chandail à manches longues et qu'elle est de 95% lorsque la personne porte un chandail à manches courtes. Ceci revient au fait que les contours ne sont pas bien définis dans certaines trames ce qui induit notre système en erreur pour le choix de la région qui contient la main dans le bras.

En général, on peut dire que nos algorithmes de détection et de suivi du visage et des mains ont une efficacité de plus de 96% sur l'ensemble des séquences testées et que cette efficacité permet la détection des différents scénarios de la prise de médicaments. Dans la section qui suit, on va évaluer nos techniques de reconnaissance de l'activité humaine qui est dans notre cas la prise de médicaments

.

Tableau 3.6 Comparaison de l'efficacité de la détection et du suivi des parties du corps pour 10 séquences testées avec des personnes qui portent des chandails à manches courtes et longues.

	Nombre de trames dans la séquence	Trames avec suivi correct du visage	Trames avec suivi correct des mains
Séquences avec des personnes qui portent des chandails à manches courtes			
Séquence François	110	107	100
Séquence Soufiane2	694	691	674
Séquence Soufiane3	321	311	300
Séquence karim1	140	137	130
Séquence Ali1	131	131	125
Séquence Younes1	237	230	220
Total	1633	1607	1549
Efficacité (%)		**98%**	**95%**
Séquences avec des personnes qui portent des chandails à manches longues			
Séquence Soufiane1	154	150	152
Séquence Atousa	145	143	144
Séquence Karim2	229	229	229
Séquence Ali2	376	376	376
Total	904	898	901
Efficacité (%)		**99%**	**99%**

3.3 La reconnaissance de l'activité humaine

3.3.1 Méthodologie expérimentale

Pour évaluer la performance de la reconnaissance de l'activité humaine, nous avons utilisé 12 séquences vidéo prises au LITIV qui représentent une activité de prise de

médicaments. Nous avons ensuite appliqué nos algorithmes sur ces séquences et on a vérifié si notre réseau de Petri est parcouru correctement. Le système a été testé sur des séquences avec une prise d'un seul médicament, de deux médicaments et aussi de trois médicament. Le système était capable de détecter avec succès la production de tous les événements dans la pluparts des séquences vidéo qu'on lui a présenté. Dans une séquence d'entrée, une détection de la prise de médicaments est correcte si notre système est capable de détecter correctement tous les états d'occlusions (états E1 à E10 du réseau de Petri, section 2.8) entre les objets d'intérêt dans la séquence et d'identifier chaque médicament pris dans cette séquence. L'efficacité de notre algorithme pour la détection de l'activité humaine a été mesurée selon la métrique suivante : *Efficacité=(Total des séquences avec détection correcte de la prise des médicaments / Total des séquences testées).*

3.3.2 Résultats

La figure 3.7 montre un graphique qui représente les périodes où les événements principaux dans la séquence Atousa ont lieu ainsi que la série de trames enregistrée par le système représentant la durée de chaque événement détecté dans cette vidéo.

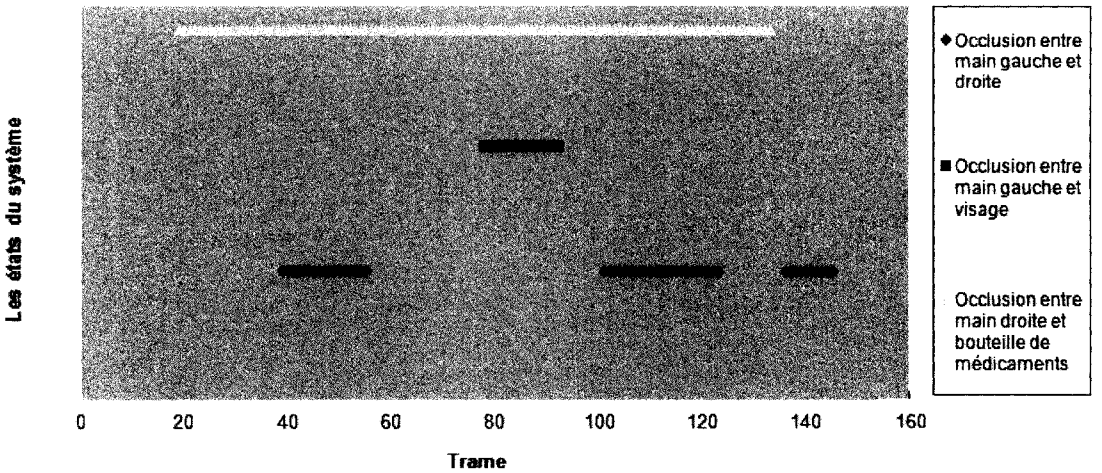

Figure 3.7 États détectés dans la séquence Atousa.

Selon la figure 3.7, initialement (première trame de la séquence vidéo), le jeton est placé à l'état Pi de notre réseau de Petri présenté à la section 2.8. Une fois que l'événement E2

(*Occlusion entre main droite et bouteille de médicament*) se produit et dure plus que 5 trames, le jeton est déplacé à l'état P2. Après, notre système détecte une occlusion entre les deux mains. La présence de cet événement alors que la bouteille est toujours manipulée par l'usager pour une durée qui excède 10 trames permet la détection de l'événement *ouverture de la bouteille de médicaments* et envoie le jeton à l'état P3. Par la suite, on remarque la détection de l'événement E4 (*Occlusion entre main gauche et visage*) qui dure plus que 10 trames et qui envoie le jeton à l'état P4. Finalement, notre système détecte le dépôt de la bouteille de médicaments sur la table et par conséquent le jeton est placé à l'état Pf et la prise de médicaments est détectée. On a évalué l'efficacité de notre système pour la reconnaissance de la prise de médicaments. Le résultat sur l'ensemble des séquences vidéo testées est présenté au tableau 3.7.

Tableau 3.7 Résultats de la détection de la prise de médicaments.

	Nombre de séquences testées	Séquences avec correct détection de la prise de médicaments
Séquences prises au LITIV avec prise de médicaments	12	9
Efficacité		**75%**

Le tableau 3.7 montre que notre système possède une efficacité de 75% et qu'il y a certaines fausses prises de médicaments comme dans le cas de la séquence Karim1. En effet, dans le traitement des trames de cette séquence, le système s'est trompé de localisation de la main dans le bras juste avant un contact entre les deux mains. La comparaison des distances entre les centroïdes des régions de la peau a détecté une occlusion entre une main et le visage alors que c'était seulement une occlusion entre les deux mains. Par conséquent, le système a détecté une prise de médicaments alors que l'usager n'a pas encore mis le comprimé dans sa bouche. Donc, lorsque le système se trompe de localisation d'une partie du corps dans la trame qui précède une occlusion, il y

a de fortes chances que le système se trompe d'événements détectés à cause de la position du centroïde de la région pour laquelle il s'est trompé de localisation.

Un autre problème survient lorsque la personne porte par exemple une chemise ouverte comme illustré à la figure 3.8 (séquence Younes2).

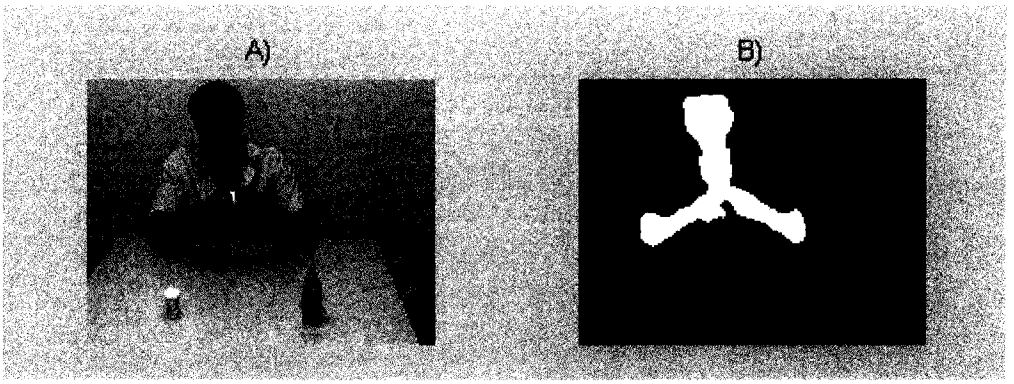

Figure 3.8 Exemple d'extraction des régions de la peau. A), Image source, B), Détection des régions de la peau contenues dans l'image source selon l'algorithme de seuillage et de segmentation présenté à la section 2.3.

L'une des limites de notre système c'est qu'il ne distingue pas le cou du visage. Pour lui, les deux constituent la région représentant le visage. Dans des cas comme celui de la figure 3.8, notre système détecte une fausse occlusion entre le visage et les mains ce qui induit le système en fausse détection de la prise des médicaments. La prochaine section présente le temps de traitement des différentes parties de notre système.

3.4 Temps d'exécution

Pour le temps d'exécution du programme, notre application ne s'exécute pas en temps réel. Les temps de traitement de différentes parties du processus de détection et de suivi ainsi que de la reconnaissance de l'activité humaine sont présentés au tableau 3.8. Ces temps ont été calculés pour une séquence de 145 trames et sur un processeur Intel Xeon™ avec une fréquence de 3.40 GHZ. Le temps de traitement est quasiment le même pour toutes les séquences testées et la seule différence se situe au niveau de la lecture de la séquence qui dépend du nombre de trames de cette dernière.

Tableau 3.8 Temps de traitement typique de différentes parties de notre système.

Partie	Durée par trame
Initialisation du système et lecture de la séquence	100 ms (varie selon le nombre de trame de la séquence)
Parcourir l'image et classifier les pixels de la peau	130 ms
Filtrage et segmentation	80 ms
Suivi du visage et des mains et mise à jour des états du réseau de Petri	600 ms
Localisation et identification des bouteilles de médicaments et mise à jour des états du réseau de Petri	850 ms
Total	1760 ms

La complexité du système est au niveau du calcul des moments de Hu d'ordre 2 pour le suivi du visage et l'identification des bouteilles de médicaments. En effet, ce calcul nécessite des boucles imbriquées ce qui augmente le temps de traitement. Aussi la comparaison des histogrammes de couleurs avec la distance MDPA qui nécessite des boucles augmentant ainsi la durée de traitement pour l'identification des bouteilles de médicaments.

CONCLUSION ET TRAVAUX FUTURS

Notre projet consistait tout d'abord à développer et à tester de nouvelles méthodes pour la localisation et le suivi du visage et des mains à partir de séquences extraites d'une seule caméra couleur statique. On a pu développer un système qui commence par détecter les régions de peau contenues dans chacune des trames. Par la suite, on utilise des hypothèses sur la forme du visage afin de localiser ce dernier dans la trame initiale. Le suivi du visage dans les trames suivantes s'effectue en utilisant les moments de Hu d'ordre 2. La localisation et le suivi des mains s'effectuent en exploitant les propriétés de contours (filtre Canny + dilatation) et du centroïde. L'identification et la localisation des bouteilles de médicament se fait en combinant les histogrammes de couleurs et les moments de Hu d'ordre 2 et le suivi de ces derniers se base sur la propriété du centroïde. Nos algorithmes de localisation et de suivi des parties du corps et des bouteilles de médicaments sont appliqués pour la détection de l'activité humaine. Dans notre cas, on a utilisé un réseau de Petri afin de reconnaître la prise de médicaments en définissant différents états liés à l'action de prise de médicaments. L'approche par apparence utilisant la forme et la couleur développée dans le cadre de notre recherche a montré son efficacité de localisation et de suivi par rapport aux approches existantes. Il est important de mentionner que nos algorithmes de détection et de suivi du visage et des mains présentés dans ce document ne requièrent aucune étape d'extraction d'arrière-plan et aucun apprentissage spécifique à l'utilisateur. En général, nos algorithmes de détection et de suivi du visage et des mains ont une efficacité de plus de 96% sur l'ensemble des séquences testées et cette efficacité permet la détection des différents scénarios de la prise de médicaments. Compte tenu de ces résultats, ces méthodes se sont montrées efficaces pour la détection de prise de médicaments et prometteurs pour être appliquées à d'autres activités humaines à domicile ou au bureau comme par exemple détecter combien de fois un employé boit du café par jour. Le système pourrait aussi servir à des fins de sécurité, afin de détecter des séquences d'événements à risques dans les endroits publics, soit par exemple un bagage abandonné.

Notre principale contribution est la réalisation d'un nouveau système de détection de l'activité humaine qui est dans notre cas la prise de médicaments. Spécifiquement, à travers notre recherche on a pu :

- créer un nouvel algorithme de suivi du visage se basant sur la détection des régions de la peau, de quelques hypothèses sur la forme du visage pour pouvoir le localiser dans la trame initiale et les moments de Hu d'ordre 2 pour effectuer le suivi. L'algorithme de détection des régions de peau a montré une efficacité de plus de 99% sur l'ensemble des images testées. Celui de la détection et du suivi du visage a montré une efficacité de plus de 98% sur l'ensemble des séquences testées.
- utiliser la segmentation en région de peau et exploiter les propriétés de contours, de morphologie mathématique (dilatation), de la densité des arrêtes ainsi que du centroïde des régions pour pouvoir localiser les mains dans le bras et les suivre. Cet algorithme de détection et du suivi des mains a montré une efficacité de plus de 96% sur l'ensemble des séquences vidéo testées.
- développer une méthode qui combine les histogrammes de couleurs et les moments de Hu pour l'identification des bouteilles de médicaments. En utilisant la table comme arrière-plan, notre algorithme a été capable en tout temps d'identifier et localiser les bouteilles de médicamentent.
- concevoir un nouveau réseau de Petri afin de reconnaitre l'activité humaine. Ce réseau a permis la reconnaissance de la prise de médicament avec une efficacité de 75%.

Travaux futurs

Pour les travaux futurs un positionnement en trois dimensions en utilisant de la stéréoscopie par exemple, pourrait permettre de réduire le nombre de faux contacts entre les objets permettant ainsi une plus efficace gestion des collisions et occlusions. Un algorithme de localisation des bouteilles de médicaments qui permettrait un suivi efficace malgré la présence d'autres objets sur la table et les occlusions avec les mains

augmentera la certitude des décisions prises. Aussi, un algorithme efficace pour la localisation de la bouche dans le visage permettant la détection du contact main-bouche plutôt que main-visage va augmenter la certitude des décisions prises. On peut aussi ajouter au système un algorithme d'identification de personnes afin de pouvoir reconnaitre qui est en train de prendre les médicaments. Donc, pouvoir identifier ce dernier à partir d'une base de données qui contiendra quelques images de différentes personnes.

RÉFÉRENCES

[1] *http://www.portailtelesante.org/article.php?sid=1327*. Visité le 04 Février 2007.

[2] Lindsay, Colin. Un portrait des aînés au Canada. *Statistique Canada*, Ottawa, octobre 1999.

[3] D. Batz, M. Batz, N. da Vitoria Lobo and M. Shah. A computer vision system for monitoring medication intake, *in Proc. IEEE 2nd Canadian Conf. on Computer and Robot Vision*, Victoria, BC, Canada, 2005, pp. 362.369.

[4] M. Valin, J. Meunier, A. St-Arnaud and J. Rousseau. Video Surveillance of Medication Intake,. *Int. Conf. of the IEEE Engineering In Medicine and Biology Society*, New York City, USA, Aug. 2006.

[5] Rein-Lien Hsu, Mohammed Abdel-Mottaleb, and Anil K. Jain. Face Detection In Color Images. *IEEE TPAMI*, Vol. 24, No. 5, pp. 696-706, May 2002.

[6] N. Eveno, "Segmentation des lèvres par un modèle déformable analytique", *Thèse de doctorat de l'INPG*, Grenoble, novembre 2003.

[7] S. Birchfield. Elliptical Head Tracking Using Intensity Gradients and Color Histograms. *Dans IEEE Proc. Computer Vision and Pattern Recognition*, pp. 232-237, 1998.

[8] N. Habili, C. Lim et A. Moini. Hand and Face Segmentation Using motion and Color Cues in Digital Image Sequences. Dans *IEEE International Conference on multimedia and Expo*, pp. 377-380, 2001.

[9] S. Hongeng et al. Video-based event recognition: activity representation and probabilistic recognition methods. Dans *Computer Vision and Image Understanding*, vol.96, no 2, pp. 129-162, novembre 2004.

[10] Peer, P., Kovac, J., and Solina, F. 2003. Human skin colour clustering for face detection. In submitted to EUROCON 2003 – *International Conference on Computer as a Tool.*

[11] Lamiaa Mostafa and Sherif Abdelazeem. Face Detection Based on Skin Color Using Neural Networks. *GVIP 05 Conference*, 19-21 December 2005, CICC, Cairo, Egypt

[12] J. Yang et A. Weibel, "A Real-Time Face Tracker", Proc. *Third Workshop Applications of Computer Vision*, pp.142-147, 1996

[13] D. Chai et A. Bouzerdoum, "A Bayesian Approach to Skin Color Classification in YCbCr Color Space", *IEEE Region Ten Conference* (TENCON'2000), vol. II, pp.421-424, Sep., 2000.

[14] Marc Kunze, Aude Billard et Sylvain Calinon : Développement d'un module de reconnaissance de mouvement et de contrôle d'imitation pour le robot humanoïde Robota. *École Polytechnique Fédérale de Lausanne*, Février 2003.

[15] S.L. Phung, A. Bouzerdoum et D. Chai. A Novel Skin Color Model in YCbCr Color Space and its Application to Human Face Detection. *Dans IEEE International Conference on Image Processing*, pp. 289-292, 2002.

[16] Karin Sobottka et Ioannis Pitas : Extraction of facial regions and features using color and shape information. *ICIP*, August 1996.

[17] Pierre-Luc Bacon, Seuillage neuronal pour la détection de peau dans une image HSV. Shawinigan, 24 février 2006.
http://pierreluc.aqra.ca/projet/bacon_pierreluc_hsv_neur.pdf

[18] Lei Xu et Erkki Oja : Randomized hough transform (RHT) : Basic mechanisms, algorithms, and computational complexities. *Dans CVGIP* : Image Understanding, volume 57(2), pages 131–154, 1993.

[19] Alexandre Lemieux : "Système d'identification de personnes par vision numérique", *Faculté des sciences et de génie*, Université Laval, Décembre 2003, pp. 30-36

[20] Ming-Hsuan Yang, David J. Kriegman et Narendra Ahuja : Detecting faces in images : A survey. Dans *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24(1), pages 34–58, 2002.

[21] Viola Jones 2001: Viola, Jones, Rapid object detection using a boosted cascade of simple features, *CVPR* 2001.

[22] Freund, Y. and Schapire, R.E. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Eurocolt 95*, Springer-Verlag, pp. 23–37.

[23] Henry A. Rowley, Shumeet Baluja et Takeo Kanade : Human face detection in visual scenes. Dans David S. Touretzky, Michael C. Mozer et Michael E. Hasselmo, éditeurs : *Advances in Neural Information Processing Systems*, volume 8, pages 875–881. The MIT Press, 1996.

[24] S. H. Cha, S. N. Srihari, "On measuring the distance between histograms", *Pattern Recognition*, Vol 35, no 6, pp 1355-1370, June 2002.

[25] Y. Rubner, C. Tomasi, L.J. Guibas, "The Earth Mover's Distance as a metric for image retrieval", *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[26] Linda G. Shapiro & George C. Stockman, *Computer Vision*, Upper Saddle River, New Jersey 07458 page 215-216.

[27] D. A. Forsyth, J. Ponce, Chap. 9 Texture, 2003. *Computer vision a modern approach*, pp.189-196. Prentice Hall.

[28] R.M. Haralick, K. Shanmugan, I. Dinstein, Textural Features for Image Classification, *IEEE Trans. On Systemes, Man, and Cybernetics*, Vol. SMC-3, No. 6, Novembre 1973. Pp. 610-621

[29] M. K. Hu. Visual pattern recognition by moment invariants. *IEEE Transactions Information Theory*, 8:179-187, 1962

[30] Ning Song Peng, Jie Yang, Zhi Liu. Mean shift blob tracking with kernel histogram filtering and hypothesis testing. *Pattern Recognition Letters* (2005) pp. 605-614.

[31] D. Comaniciu, V. Ramesh, P. Meer, Real-Time Tracking of Non-Rigid Objects using Mean Shift, BEST PAPER AWARD, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR00)*, Hilton Head Island, South Carolina, Juin 1997. Vol. 2, 142-149.

[32] Mathias FONTMARTY. Suivi 3D de mouvements humains pour l'interaction homme-robot. *EDSys 2007* – 8e congrès des doctorants

[33] A. Naeem, S. Mills, and T. Pridmore, Structured Combination of Particle Filter and Kernel Mean-Shift Tracking, *Proc. Int. Conf. Image and Vision Computing*, New Zealand, 2006.

[34] A. Jacquot, P. Sturm, O. Ruch, "Adaptative Tracking of Non-Rigid Objects Based on Color Histograms and Automatic Parameter Selection", *IEEE Workshop on Motion and Video Computing*, pp 103-109, Breckenridge, Colorado, January 2005.

[35] Mun-Ho Jeong, Bum-Jae You, Yonghwan Oh, and Sang-Rok Oh. Adaptive Mean-Shift Tracking with Novel Color Model. *Proceedings of the IEEE International Conference on Mechatronics & Automation Niagara Falls*, Canada • July 2005.

[36] J. Gao, A. G. Hauptman, A. Bharucha et H. D. Wactlar. Dining activity Analysis Using a Hidden Markov Model. Dans Proc. Of *the 17th International Conference on Pattern Recognition*, Cambridge, Royaume-Uni, vol.2, pp.915-918, 2004.

[37] Fuentes, L.M. and Velastin, S.A., 2001, "People tracking in surveillance applications", 2nd *IEEE International Workshop on Performance Evaluation on Tracking and Surveillance*, PETS, Kauai (Hawaii-USA), 14/12/2001

[38] N. Rota et M. Thonnat, Activity Recognition from Video Sequences using Declarative Models. Dans 14th. *European Conference on Artificial Intelligence 2000, Berlin*, Allemagne, pp. 673-680, 2000.

[39] N. Ghanem et. al, "Representation and recognition of events in surveillance video using petri nets," in *Event Detection and Recognition Workshop at ICCV*. IEEE, June 2004.

[40] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Systems, Man, and Cybernetics,* Vol. 9, No. 1, pp. 62-66, 1979.

[41] P. F. Gabriel, J. G. Verly, J. H. Piater, and A. Genon. The state of the art in multiple object tracking under occlusion in video sequences. In Proc. *ACIVS,* 2003.

[42] S. Carbini, J.E. Viallet, O. Bernier, "Simultaneous Body Parts Statistical Tracking for Bi-Manual Interactions", *ORASIS*, Fournol, France, 24-27 may 2005.

[43] A. Choksuriwong H. Laurent B. Emile. Comparative Study of Objects Invariant Descriptors. *ENSI de Bourges* - Universite d'Orleans - *Laboratoire Vision et Robotique* - UPRES EA 2078 10 boulevard Lahitolle, 18020 Bourges Cedex, France.

[44] Canny, J.F. "A Computational Approach to Edge Detection". IEEE Transactions on Pattern Analysis and Machine Intelligence. 1986,8(6) pp.679-698.

[45] N. Ghanem et. al, "Representation and recognition of events in surveillance video using Petri nets," in *Event Detection and Recognition Workshop at ICCV*. IEEE, June 2004.

[46] S. Ammouri, G-A. Bilodeau. Face and hands detection and tracking applied to the monitoring of medication intake. In *Canadian Conference on Computer and Robot Vision*, Windsor, Canada, pp. 147-154, 2008.

ANNEXE I

Article publié à CRV

L'article présenté ici [46] a été publié lors de la 5^{e} conférence canadienne *Computer and Robot Vision* d'IEEE qui s'est tenue à Windsor du 28 au 30 Mai 2008 et a fait l'objet d'une présentation orale.